Journal of Computer and System Sciences **56**, 223–252 (1998)
Article No. SS971560

Covering Cubes by Random Half Cubes, with Applications to Binary Neural Networks*

Jeong Han Kim

Microsoft Research, One Microsoft Way, Redmond, Washington 98052
E-mail: jehkim@microsoft.com

and

James R. Roche

Center for Communications Research, Thanet Road, Princeton, New Jersey 08540
E-mail: roche@ccr-p.ida.org

Received October 27, 1995; revised October 24, 1997

Let Q_n be the (hyper)cube $\{-1, 1\}^n$. This paper is concerned with the following question: How many vectors must be chosen uniformly and independently at random from Q_n before every vector in Q_n itself has negative inner product with at least one of the random vectors? For any fixed $\epsilon > 0$, a simple expectation argument shows that for all sufficiently large n, $(1+\epsilon)\, n$ random vectors suffice with high probability. In this paper we prove that there are $\epsilon, \rho > 0$ such that $(1-\epsilon)\, n$ random vectors are also enough and such that at least ρn random vectors are necessary.

This problem is partially motivated by neural network problems. Neural networks are being used to solve a growing number of difficult problems such as speech recognition, handwriting recognition, and protein structure prediction. Recently, for both theoretical and practical reasons, neural networks with binary weights (binary neural networks) have attracted much attention. In spite of considerable analysis based on statistical mechanics, the following two basic questions about binary neural networks have remained unanswered.

Q1. Is there a positive constant ρ such that for all sufficiently large n there is a binary neural network of n neurons which can separate ρn (unbiased) random patterns with probability close to 1?

Q2. Is it possible for a binary neural network of n neurons to separate $(1-o(1))\, n$ random patterns with probability greater than some positive constant? (Here $o(1)$ goes to 0 as n goes to infinity.) This question is raised because no binary neural network of n neurons can separate more than n random patterns with probability close to 1.

Our results yield the answers "YES" to Q1 and "NO" to Q2.

1. INTRODUCTION

Let Q_n be the (hyper)cube $\{-1, 1\}^n$. Let the half cube H_x generated by any (n-dimensional) real vector x be the set of all real vectors w having negative inner product with x, that is,

$$H_x := \{w \in \mathbb{R}^n : w \cdot x < 0\}.$$

A half cube generated by x is $H_x \cap Q_n$. If $w \in H_x$, it is natural to say that H_x *covers* the vector w. With some abuse of notation, we also say in this case that the vector x covers w. A *random vector* is to be uniformly chosen at random from all vectors in Q_n. This paper is concerned with the following question: How many half cubes generated by independent random vectors from Q_n are needed to cover every vector in Q_n at least once? In other words, how many random vectors from Q_n are needed to cover all of Q_n?

For k random vectors $X^{(1)}, ..., X^{(k)}$ in Q_n, $P_{b,b}(n, k)$ denotes the probability (with the subscript b for "binary") that the random vectors do not cover Q_n. That is,

$$P_{b,b}(n, k) := \Pr(\exists z \in Q_n \text{ s.t. } X^{(j)} \cdot z \geqslant 0 \text{ for all } j = 1, ..., k)$$
$$= \Pr\left(Q_n \not\subseteq \bigcup_{j=1}^{k} H_{X^{(j)}}\right).$$

A straightforward expectation argument shows that for any positive $\varepsilon > 0$,

$$P_{b,b}(n, (1+\varepsilon)\, n) \to 0 \qquad \text{as} \quad n \to \infty.$$

(Of course, $(1+\varepsilon)\, n$ actually means $\lfloor (1+\varepsilon)\, n \rfloor$. For the sake of simplicity, floor or ceiling signs will be omitted.) A previously unresolved question is whether there exists $\varepsilon > 0$ so that the probability $P_{b,b}(n, (1-\varepsilon)\, n)$ still goes to 0. Another heretofore open question is whether there is a

* This work was carried out while both authors were at AT&T Bell Laboratories.

0022-0000/98 $25.00

constant $\rho > 0$ such that $P_{b,b}(n, \rho n) \to 1$. It is still unknown whether the following conjecture (from [10]) is true.

Conjecture 1.1. There exists a positive constant $c_{b,b}$ such that

$$P_{b,b}(n, cn) \to 1 \quad \text{if} \quad c < c_{b,b} \quad \text{and}$$
$$P_{b,b}(n, cn) \to 0 \quad \text{if} \quad c > c_{b,b}.$$

Finding good bounds on $c_{b,b}$, if it exists, is a problem which has attracted many researchers (see, e.g. [25, 26]). There are simulation and non-rigorous analytical results. Krauth and Opper's simulation result [26] suggests that $c_{b,b}$ is approximately 0.82. Because the simulation result is based on experiments with n less than 25, though, the extrapolation to arbitrarily large n is somewhat speculative. Krauth and Mézard [25] obtained $c_{b,b} \approx 0.83$ using the replica method (from statistical mechanics) with one so-called symmetry breaking; however, they did not give rigorous arguments. Therefore, in spite of the results mentioned above, it was unknown whether there is a positive constant ε such that

$$P_{b,b}(n, (1-\varepsilon)\, n) \to 0 \quad \text{as} \quad n \to \infty.$$

Concerning a lower bound, no rigorous approach had been devised to prove that there is a positive constant ρ such that

$$P_{b,b}(n, \rho n) \to 1 \quad \text{as} \quad n \to \infty.$$

In this paper, we demonstrate that positive ε and ρ can be found. The existence of a sharp threshold $c_{b,b}$ remains a conjecture.

THEOREM 1.2. *For all* $\varepsilon < 0.0037$,

$$\lim_{n \to \infty} P_{b,b}(n, (1-\varepsilon)\, n) = 0.$$

THEOREM 1.3. *For all* $\rho < 0.005$,

$$\lim_{n \to \infty} P_{b,b}(n, \rho n) = 1.$$

We will actually prove a slightly stronger theorem than Theorem 1.3 to answer a neural network question.

THEOREM 1.4. *For all* $\rho < 0.005$,

$$\lim_{n \to \infty} n(1 - P_{b,b}(n, \rho n)) = 0.$$

The constant 0.005 in Theorem 1.4 (which will yield the lower bound in Theorem 3.1) can be improved in at least two different ways. Simply by doing a more painstaking analysis of the scheme that we consider, we believe that one can raise the lower bound to 0.03. A similar but more sophisticated scheme suggested to the second author by Paul Lemke appears to yield a lower bound of approximately 0.3, but the work required to make this bound completely rigorous seems daunting.

In this paper we draw all vectors from the cube Q_n. More generally, one can choose either the covering vectors or the vectors to be covered (or both) from the unit sphere S_n of $\mathbb{R}^n$. For example, one might cover either Q_n or S_n with random vectors from S_n (chosen with respect to normalized Lebesgue measure, of course). A more general setting will be discussed in Section 3.

The problem of covering the unit sphere was considered in the late 1950s and early 1960s for geometrical reasons, and Wendel [37] (see also [5]) found the exact formula for the corresponding probabilities $P_{s,s}(n, k)$. Researchers had actually been interested in finding the probability that k random vectors in S_n lie on some hemisphere, which is equivalent to our covering problem. The formula

$$P_{s,s}(n, k) = 2^{-k+1} \sum_{i=0}^{n-1} \binom{k-1}{i} \tag{1.1}$$

was proved based on Schläfli's result on the number of linearly separable dichotomies of points in general position. (Wendel proved that the result is true for more general probability measures on S_n.) This formula implies, in particular, that at least $(2-\varepsilon)\, n$ random vectors are necessary and $(2+\varepsilon)\, n$ vectors are sufficient.

Erdős asked how many random binary vectors (i.e., random vectors from Q_n) are needed to cover S_n. Füredi [9] proved that the corresponding probabilities $P_{b,s}(n, k)$ are almost the same as $P_{s,s}(n, k)$. (See also Venkatesh [36]. Earlier Komlós (unpublished) found upper and lower bounds for the probability. For earlier work see Komlós [23].) Namely,

$$P_{b,s}(n, k) = 2^{-k+1} \sum_{i=0}^{n-1} \binom{k-1}{i} + O\left(\frac{1}{\sqrt{n}}\right). \tag{1.2}$$

The proof uses the fact—established by Komlós [23]—that the probability of a random $n \times n$ binary (i.e. ± 1) matrix being singular is $O(1/\sqrt{n})$. (See also [3].) A recent result of Kahn *et al.* [19] proves that the singularity probability is $O((0.999)^n)$, which yields

$$P_{b,s}(n, k) = 2^{-k+1} \sum_{i=0}^{n-1} \binom{k-1}{i} + O((0.999)^n).$$

These results imply that when covering S_n by random binary vectors, at least $(2-\varepsilon)\, n$ random vectors are again

necessary and $(2+\varepsilon)n$ vectors are again sufficient. The problem of covering Q_n turns out to be harder and has not yet been settled.

In the next two sections, which are independent of the other sections, we briefly introduce neural networks and give definitions of separating capacity in a general framework. In the last two sections, we prove Theorems 1.2 and 1.4.

2. A MOTIVATION: NEURAL NETWORKS

Neural networks are being used to solve a growing number of difficult problems such as speech recognition, handwriting recognition, and protein structure prediction. A number of researchers have also considered using such networks to store information as the brain does. Current neural network models are based on the following three natural assumptions. (For history and related work see [17, 21, 22]. See also [28] and [29] for an earlier model.)

A1. Neural network models, or simply neural networks, are interconnected systems of neurons with binary activity (see [31]).

A2. Interconnections among the neurons collectively encode information (cf. [27]).

A3. A neural network evolves using certain weights, called synaptic weights or learning rules.

Neural networks with binary weights (binary neural networks) have attracted much attention [2, 13–15, 20, 30, 35] for both theoretical and practical reasons. From a theoretical standpoint, the discrete structure of a binary neural network has very interesting properties [14, 33]. From a practical standpoint, we clearly do not want to require infinite precision for the weights, because we wish to encode them using a modest number of bits.

In spite of considerable analysis based on statistical mechanics [14, 30, 33], the basic questions regarding binary neural networks have remained largely unanswered. Even the following questions, which are almost the same as the first two questions raised (and answered) in the previous section, have remained open.

Q1. Is there a positive constant ρ such that for all sufficiently large n there is a binary neural network of n neurons which can separate ρn unbiased random patterns with probability close to 1?

Q2. Is it possible for a binary neural network of n neurons to separate $(1-o(1))n$ random patterns with probability greater than some positive constant? (Here $o(1)$ goes to 0 as n goes to infinity.) This question is raised because no binary neural network of n neurons can separate more than n random patterns with probability close to 1. (See, e.g., [12] for details.)

Theorems 1.2 and 1.4 answer Q1 and Q2: "YES" to Q1 and "NO" to Q2. Our approaches are combinatorial and completely rigorous.

By introducing an energy function, Hopfield [18] provided useful physical insight into A1–A3. His model consists of a system of fully interconnected neurons with interconnections having certain (non-binary) weights. The weights are now well known as Hebb's rule (cf. [16]). It was Gardner [10, 11] who considered neural network models in a more general framework. She raised the question of determining the optimal separating capacity (roughly speaking, the maximum ρ in Q1) of all neural networks with certain constraints on their weights, in particular, of binary neural networks. Gardner and Derrida [12] used the non-rigorous replica method to give answers for spherical neural networks, which have weight vectors drawn from the unit sphere. They also noted that the replica method gave a clearly incorrect answer for binary neural networks.

Few rigorous results regarding binary neural networks have been obtained. One notable exception is Venkatesh's theorem [35] stating that the capacity function (roughly speaking, the maximum ρ as a function of n in Q1) of the clipped Hebb's rule, or majority rule, is $1/(\pi \log n)$ (see also [13]). Thus its capacity is 0. Venkatesh also introduced another network having capacity function of order at least $1/\log n$.

The positive answer to Q1 follows from an explicit multi-stage construction generalizing the clipped Hebb's rule analyzed by Venkatesh. The construction is simple to describe, and the capacity function of the corresponding neural network lends itself readily to heuristic analysis. Some effort is required, though, to make the heuristic analysis precise. The negative answer to Q2 is based on the fact that the optimal binary weights cannot be far from the weights determined by Hebb's rule, where the distance between an arbitrary real vector $v=(v_i)$ and a binary (± 1) vector $w=(w_i)$ is $\sum_i |v_i| - \sum_i w_i v_i$.

3. SEPARATING CAPACITY: A GENERAL SETTING

This sections gives definitions related to separating capacity of random patterns in a general framework.

First of all, in a theoretical definition of a neural network we do not need the fact that a neuron has binary activity. Hence A1 may be replaced by the weaker statement below.

A1′. Neural network models are interconnected systems of neurons.

Let $\mathscr{M}_n$ be the set of all $n \times n$ real matrices with all diagonal entries 0. For $J=(J_{ij}) \in \mathscr{M}_n$, $J^{(r)}$ denotes the rth row vector without the diagonal entry J_{rr}, i.e.,

$$J^{(r)} := (J_{r1}, ..., J_{rr-1}, J_{rr+1}, ..., J_{rn}).$$

For $W_{n-1} \subseteq \mathbb{R}^{n-1}$, $\mathcal{M}_n(W_{n-1})$ denotes the collection of all matrices whose row vectors without the diagonal entries are in W_{n-1}.

A pair $\mathcal{N}_n = (\Omega_n, J_n)$ is a *neural field of n neurons with random patterns*, or simply a *neural field*, if $\Omega_n = (V_n, \mathrm{P}_n)$ is a probability space with $V_n \subseteq \mathbb{R}^n$, and J_n is a function from $\bigcup_{k=1}^{\infty} (V_n)^k$ into $\mathcal{M}_n$. The probability space Ω_n is called a *pattern space*. The matrix-valued function J_n is a *weight matrix* (or *algorithm*), whose entries are *weights*. In neural networks, a dynamic system is governed by a weight matrix, which is a function of patterns in Ω. The weight matrix plays a role similar to that played by a field (e.g., magnetic field) in a physical system. Though only random patterns are considered in this paper, patterns can also be chosen arbitrarily, in which case V_n is a set of patterns without any associated probability measure.

If a neural field $\mathcal{N}_n = (\Omega_n = (V_n, \mathrm{P}_n), J_n)$ has $J_n \in \mathcal{M}_n(W_{n-1})$, i.e., if the image of J_n is a subset of $\mathcal{M}_n(W_{n-1})$, then it is called a (Ω_n, W_{n-1})-neural field or, more broadly, a (V_n, W_{n-1})-neural field regardless of the probability measure. A *binary neural field of binary neurons* is a (Q_n, Q_{n-1})-neural field, and a *spherical neural field of binary neurons* is a (Q_n, S_{n-1})-neural field, where S_{n-1} is the unit sphere in $\mathbb{R}^{n-1}$. Binary and spherical neural fields of spherical neurons can be defined similarly. A *pattern of n neurons*, or simply a *pattern*, is a vector $X = (X_i)$ in V_n, where X_i represents the state of neuron i. Thus one might express assumption A1 from the previous section imply as $V_n = Q_n$. A *random pattern* of a neural field $\mathcal{N}_n = (\Omega_n, J_n)$ is a random vector of Ω_n.

EXAMPLE 1 (Hopfield Model). $\Omega_n = Q_n$ with uniform distribution, and $J_n(X^{(1)}, ..., X^{(k)}) = (\sum_{t=1}^{k} X_i^{(t)} X_j^{(t)})_0$; that is, the ij-entry of J_n is

$$\sum_{t=1}^{k} X_i^{(t)} X_j^{(t)} \text{ (Hebb's Rule)} \quad \text{for } i \neq j.$$

(The subscript 0 indicates that all diagonal entries of J_n are 0.)

EXAMPLE 2 (Clipped Hebb's Rule or Majority Rule). $\Omega_n = Q_n$ with uniform distribution, and

$$J_n(X^{(1)}, ..., X^{(k)}) = \left(\operatorname{sgn} \left(\sum_{t=1}^{k} X_i^{(t)} X_j^{(t)} \right) \right)_0,$$

where

$$\operatorname{sgn}(z) := \begin{cases} 1 & \text{if } z \geqslant 0 \\ -1 & \text{if } z < 0. \end{cases}$$

Note that this is a binary neural field.

A neural field $\mathcal{N}_n = (\Omega_n, J_n)$ can *separate* k patterns $X^{(1)}, ..., X^{(k)}$ if

$$\sum_{j=1}^{n} J_{ij} X_j^{(t)} X_i^{(t)} \geqslant 0 \quad \text{for all } t = 1, ..., k \text{ and } i = 1, ..., n,$$

where J_{ij} is the ij-entry of $J_n(X^{(1)}, ..., X^{(k)})$. Let $0 < \epsilon < 1$, and let $X^{(1)}, ..., X^{(k)}$ be k mutually independent random patterns of $\mathcal{N}_n = (\Omega_n, J_n)$. The *$\epsilon$-capacity* of $\mathcal{N}_n$ is the largest number $c = c(\mathcal{N}_n; \epsilon) \in \mathbb{R} \cup \{\infty\}$ such that

$$\Pr(\mathcal{N}_n \text{ can separate } X^{(1)}, ..., X^{(k)}) \geqslant 1 - \epsilon \quad \text{for all} \quad k \leqslant cn.$$

The *ϵ-capacity function* of a sequence of neural fields $\{\mathcal{N}_n\}$ is the sequence $\{c(\mathcal{N}_n; \epsilon)\}$. (See [34] and references therein for related definitions and history.) The *inf-capacity* $c^{(-)}(\{\mathcal{N}_n\})$ and *sup-capacity* $c^{(+)}(\{\mathcal{N}_n\})$ of a sequence of neural fields $\{\mathcal{N}_n\}$ are defined by

$$c^{(-)}(\{\mathcal{N}_n\}) := \lim_{\epsilon \to 0} \liminf_{n \to \infty} c(\mathcal{N}_n; \epsilon),$$

$$c^{(+)}(\{\mathcal{N}_n\}) := \lim_{\epsilon \to 0} \limsup_{n \to \infty} c(\mathcal{N}_n; \epsilon).$$

Let $\{(\Omega_n = (V_n, \mathrm{P}_n), W_{n-1}\}_{n=1, 2, ...}$ be a sequence of pairs of probability spaces Ω_n and subsets W_{n-1} of $\mathbb{R}^{n-1}$ with $V_n \subseteq \mathbb{R}^n$. We define the optimal separating ϵ-capacity, or simply ϵ-capacity, of all (Ω_n, W_{n-1})-neural fields by

$$c(\Omega_n, W_{n-1}; \epsilon) := \sup\{c(\mathcal{N}_n; \epsilon): \mathcal{N}_n \text{ is a } (\Omega_n, W_{n-1})\text{-neutral field}\}$$

and the inf-capacity and sup-capacity of the sequences by

$$c^{(-)}(\{(\Omega_n, W_{n-1})\}) := \lim_{\varepsilon \to 0} \liminf_{n \to \infty} c(\Omega_n, W_{n-1}; \epsilon),$$

$$c^{(+)}(\{(\Omega_n, W_{n-1})\}) := \lim_{\varepsilon \to 0} \limsup_{n \to \infty} c(\Omega_n, W_{n-1}; \epsilon).$$

When $c^{(+)}(\{(\Omega_n, W_{n-1})\}) = c^{(-)}(\{(\Omega_n, W_{n-1})\})$, we write $c(\{(\Omega_n, W_{n-1})\})$ for their common value and call it the (optimal) capacity of all (Ω_n, W_{n-1})-neural fields. Regarding Q_n and S_n as probability spaces with uniform and normalized Lebesgue measures, respectively, denote

$$c_{b,b}^{(\pm)} := c^{(\pm)}(\{(Q_n, Q_{n-1})\}),$$

$$c_{b,s}^{(\pm)} := c^{(\pm)}(\{(Q_n, S_{n-1})\}),$$

$$c_{s,b}^{(\pm)} := c^{(\pm)}(\{(S_n, Q_{n-1})\}),$$

$$c_{s,s}^{(\pm)} := c^{(\pm)}(\{(S_n, S_{n-1})\}).$$

If $c_{b,b}^{(-)} = c_{b,b}^{(+)}$, one might call $c_{b,b}$ the optimal capacity of binary neural fields. Similarly, $c_{b,s}$ could be called the optimal capacity of spherical neural fields.

As remarked before, (1.1) and (1.2) yield

$$c_{s,s} = c_{b,s} = 2.$$

Theorems 1.2 and 1.4 imply the following theorem.

THEOREM 3.1. *For every $\epsilon > 0$ there exists $n_0(\epsilon)$ such that*

$$c(Q_n, Q_{n-1}; \epsilon) \geqslant 0.005 \quad \textit{for all } n \geqslant n_0(\epsilon)$$

and

$$c(Q_n, Q_{n-1}; 1-\epsilon) \leqslant 0.9963 \quad \textit{for all } n \geqslant n_0(\epsilon).$$

In particular,

$$0.005 \leqslant c_{b,b}^{(-)} \leqslant c_{b,b}^{(+)} < 0.9963.$$

4. UPPER BOUND: THE PROOF OF THEOREM 3.1

For the proofs of Theorems 1.2 and 1.4, it is convenient to think of the random vectors $X^{(1)}, ..., X^{(k)} \in Q_n$ as the rows of the random $k \times n$ matrix $X = X(n, k)$ in which all entries are equally likely to be $+1$ or -1. Hence one may regard $X^{(i)}$ as the ith row vector of X. We also write

$$X_{ij} \qquad \text{for } X_j^{(i)},$$

and write

$$Xz \geqslant 0 \qquad \text{if} \quad \forall i = 1, 2, ..., k \qquad X^{(i)} \cdot z \geqslant 0.$$

For a constant $0 < \rho < 1$, we will write $P_{b,b}(n, \rho)$ for $P_{b,b}(n, \lfloor \rho n \rfloor)$. Though this is, of course, a non-trivial abuse of notation, no confusion seems to arise.

4.1. Sketch of the Proof

This subsection presents the idea of the proof of Theorem 1.2 with a slightly worse constant than 0.0037. We use "$\approx$" to mean approximately equal and emphasize that what follows is only an aid to our intuition. In other words, we should have written in many places "informally speaking," "in some suitable sense," etc.

Let $X := X(n, 1-\varepsilon)$, $0 < \varepsilon < 1$. Define $\mathscr{A}_z = \mathscr{A}_z(n, 1-\varepsilon)$ to be the event $Xz \geqslant 0$, and let U_j be the jth column sum of X:

$$U_j := \sum_{i=1}^{k} X_{ij}.$$

These column sums were also used in [24] to solve some neural network problems. Suppose now that $\mathscr{A}_z$ occurs. Then

$$\sum_{i=1}^{k} \left| \sum_{j=1}^{n} z_j X_{ij} \right| = \sum_{i=1}^{k} \sum_{j=1}^{n} z_j X_{ij} = \sum_{j=1}^{n} \sum_{i=1}^{k} z_j X_{ij}$$
$$= \sum_{j=1}^{n} z_j \sum_{i=1}^{k} X_{ij} = \sum_{j=1}^{n} z_j U_j = z \cdot U, \qquad (4.3)$$

where $U = (U_1, ..., U_n)$. On the other hand, the distributions of

$$n^{-1/2} \sum_{j=1}^{n} z_j X_{ij} \quad \text{and} \quad k^{-1/2} U_j = k^{-1/2} \sum_{i=1}^{k} X_{ij}$$

are asymptotically standard normal because they are normalized sums of independent ± 1 random variables. One expects, therefore, that the random variables

$$Y_z^{(n)} := \sum_{i=1}^{k} \left| \sum_{j=1}^{n} z_j X_{ij} \right| \quad \text{and}$$
$$Z_n := \sum_{j=1}^{n} |U_j| = \sum_{j=1}^{n} \left| \sum_{i=1}^{k} X_{ij} \right|$$

are highly predictable, that is, they are highly concentrated near their means: for all $\lambda > 0$,

$$\Pr(|Y_z^{(n)} - \mathrm{E}[Y_z^{(n)}]| > \lambda \mathrm{E}[Y_z^{(n)}]) \leqslant e^{-a_\lambda n} \qquad (4.4)$$

and

$$\Pr(|Z_n - \mathrm{E}[Z_n]| > \lambda \mathrm{E}[Z_n]) \leqslant e^{-b_\lambda n}, \qquad (4.5)$$

where a_λ and b_λ are constants depending only on λ. (In the rigorous proof, the exact values of a_λ and b_λ are important because our choice of ε depends on these values.) Moreover,

$$\mathrm{E}[Y_n^{(n)}] = \mathrm{E}\left[\sum_{i=1}^{k} \left| \sum_{j=1}^{n} z_j X_{ij} \right| \right] = \mathrm{E}\left[\sum_{i=1}^{k} \left| \sum_{j=1}^{n} X_{ij} \right| \right]$$

since the X_{ij} are ± 1 with equal probability. This implies that as ε tends to zero (or equivalently, as k tends to n), $\mathrm{E}[Y_z^{(n)}]$ tends to $\mathrm{E}[Z_n]$, which together with (4.4) and (4.5) implies that

$$Y_z^{(n)} \approx Z_n \quad \text{for small enough } \varepsilon > 0 \text{ and sufficiently large } n. \qquad (4.6)$$

Let $\mathscr{U}$ ($= \mathscr{U}(n, \varepsilon)$) be the set of all possible vector values of U. Then

$$\mathrm{P}_{b,b}(n, 1-\varepsilon) = \Pr\left(\bigcup_{z \in Q_n} \mathscr{A}_z\right)$$

$$= \sum_{u \in \mathscr{U}} \Pr(U=u) \Pr\left(\bigcup_{z \in Q_n} \mathscr{A}_z \mid U=u\right). \quad (4.7)$$

Suppose that $\mathscr{A}_z$ occurs and that $U=u \in \mathscr{U}$. Then it is not hard to see that (4.3) and (4.6) imply

$$z \cdot u = Y_z^{(n)} \approx Z_n = \sum_{j=1}^{n} |u_j| \quad (4.8)$$

for small enough $\varepsilon > 0$ and sufficiently large n. Define $Q(u)$ to be the set of all $z \in Q_n$ satisfying (4.8), and define δ_ε by the equation $e^{\delta_\varepsilon n} = |Q(u)|$. Then

$$\Pr\left(\bigcup_{z \in Q_n} \mathscr{A}_z \mid U=u\right) \approx \Pr\left(\bigcup_{z \in Q(u)} \mathscr{A}_z \mid U=u\right)$$

$$\leqslant \sum_{z \in Q(u)} \Pr(\mathscr{A}_z \mid U=u)$$

$$\leqslant e^{\delta_\varepsilon n} \max_{z \in Q_n} \Pr(\mathscr{A}_z \mid U=u). \quad (4.9)$$

Also, $z \in Q(u)$ implies that z must be close to the signature vector $z_u := (\mathrm{sgn}(u_j))_{j=1}^n$ of u. Thus we expect to have

$$\lim_{\varepsilon \to 0} \delta_\varepsilon = 0. \quad (4.10)$$

Therefore, it is (intuitively) enough to show that there is a universal constant $\alpha > 0$ such that

$$\Pr(\mathscr{A}_z \mid U=u) \leqslant e^{-\alpha k} = e^{-\alpha(1-\varepsilon)n}$$

$$\text{for all} \quad z \in Q_n \qquad \text{and} \qquad u \in \mathscr{U}_0, \quad (4.11)$$

where $\mathscr{U}_0 \subseteq \mathscr{U}$ and $\Pr(U \in \mathscr{U}_0) \approx 1$. This is because $\delta_\varepsilon - (1-\varepsilon)\alpha \approx -\alpha$ for sufficiently small ε (see (4.10)), and (4.11) along with (4.7) and (4.9) (intuitively) implies that

$$\mathrm{P}_{b,b}(n, 1-\varepsilon)$$

$$= \sum_{u \in \mathscr{U}} \Pr(U=u) \Pr\left(\bigcup_{z \in Q_n} \mathscr{A}_z \mid U=u\right)$$

$$\leqslant \sum_{u \in \mathscr{U}_0} \Pr(U=u) \Pr\left(\bigcup_{z \in Q_n} \mathscr{A}_z \mid U=u\right) + \Pr(U \notin \mathscr{U}_0)$$

$$\lesssim \exp(n(\delta_\varepsilon - (1-\varepsilon)\alpha)).$$

The proof of (4.11) is based on the central limit theorem (see, e.g., [4]), the FKG inequality (see, e.g., [1]), and the facts that the events $\{X_{ij}=1\}_{j=1,\ldots,k}$ given $U=u$ (i fixed) are mutually independent and

$$\Pr(X_{ij}=1 \mid U=u) = \tfrac{1}{2}(1+u_j/k) \quad \text{for all } j=1, \ldots, n.$$

We will actually obtain a better upper bound in (4.11), which depends on $z \cdot u$.

4.2. Lemmas

This subsection introduces a few lemmas from which Theorem 1.2 easily follows. The next sections are for the proofs of the lemmas.

Denote, for all $\lambda \in \mathbb{R}$,

$$\Phi(\lambda) := (2\pi)^{-1/2} \int_{-\infty}^{\lambda} e^{-t^2/2}\, dt,$$

$$f(\lambda) := \log(\Phi(\lambda)),$$

$$g(\lambda) := (2\pi)^{-1/2} \int_{-\infty}^{\infty} \log(\cosh(\lambda t))\, e^{-t^2/2}\, dt$$

and

$$h(\theta; \lambda) := \theta\lambda + \lambda^2/2 + f(-\lambda) + \log 2 \quad \text{for } 0 < \theta \leqslant (2/\pi)^{1/2}.$$

(Our logarithms are base e.) Then take positive numbers ε_0, θ_0, λ_0, and μ_0 with $\mu_0 \geqslant (2/\pi)^{1/2}$ such that

$$\varepsilon_0 \log 2 + (1-\varepsilon_0)\, h(\theta_0; \lambda_0) \leqslant 0 \quad (4.12)$$

and

$$-(1-\varepsilon_0)^{1/2}\, \theta_0\mu_0 + g(\mu_0) + \log 2 + (1-\varepsilon_0)\, f(\theta_0) \leqslant 0. \quad (4.13)$$

For example, we may take $\varepsilon_0 = 0.0037$, $\theta_0 = 0.7465$, $\lambda_0 = 0.148$ and $\mu_0 = 2.325$.

Since

$$h(\theta_0; \lambda_0) < 0 \quad \text{and} \quad f(\theta_0) < 0$$

it is easy to see that, for every $0 < \varepsilon < \varepsilon_0$, there is $0 < \delta := \delta(\varepsilon) < 1$ such that

$$\varepsilon \log 2 + (1-\delta)(1-\varepsilon)\, h(\theta_0; \lambda_0) < -\delta \quad (4.14)$$

and

$$-(1-\varepsilon)^{1/2}\, \theta_0\mu_0 + (1+\delta)\, g(\mu_0)$$

$$+ \log 2 + (1-\varepsilon)\, f(\theta_0 + \delta) < -2\delta. \quad (4.15)$$

Moreover, (4.15) gives

$$-(1-\varepsilon)^{1/2}\,\theta\mu_0+(1+\delta)\,g(\mu_0)+\log 2 \\ +(1-\varepsilon)\,f(\theta+\delta)<-2\delta \qquad \text{for all} \quad \theta\geqslant\theta_0 \quad (4.16)$$

because the partial derivative of the left side with respect to θ is always negative. (Recall that $\mu_0\geqslant(2/\pi)^{1/2}$.) In what follows, we assume that ε and δ are fixed and satisfy the above inequalities.

Let $\mathscr{B}_z$ be the event

$$\mathscr{B}_z:=\left\{\sum_{i=1}^{k}\left|\sum_{j=1}^{k}z_jX_{ij}\right|\geqslant\theta_0 n^{1/2}(1-\varepsilon)\,n,\ z\in Q_n\right\} \quad \text{(c.f. (4.4))}$$

(recall that $k:=\lfloor(1-\varepsilon)\,n\rfloor$), and define

$$\mathscr{U}_1:=\left\{u\in\mathscr{U}\colon\sum_{j=1}^{n}|u_j|\leqslant(1+\delta)(2k/\pi)^{1/2}\,n\right\} \quad \text{(c.f. (4.5))},$$

$$\mathscr{U}_2:=\left\{u\in\mathscr{U}\colon\sum_{j=1}^{n}\log(\cosh(\mu_0k^{-1/2}u_j))\leqslant(1+\delta)\,g(\mu_0)\,n\right\},$$

and $\mathscr{U}_0:=\mathscr{U}_1\cup\mathscr{U}_2$.

Note that

$$\begin{aligned}\mathrm{P}_{b,b}(n,1-\varepsilon)&=\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_z\right)\\&\leqslant\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_n\cap\mathscr{B}_z\right)+\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_z\cap\bar{\mathscr{B}}_z\right)\\&\leqslant\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_z\cap\mathscr{B}_z\right)+\sum_{z\in Q_n}\Pr(\mathscr{A}_z\cap\bar{\mathscr{B}}_z),\end{aligned} \quad (4.17)$$

and

$$\begin{aligned}&\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_z\cap\mathscr{B}_z\right)\\&\quad=\sum_{u\in\mathscr{U}}\Pr(U=u)\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_z\cap\mathscr{B}_z\mid U=u\right)\\&\quad\leqslant\Pr(U\notin\mathscr{U}_0)+\sum_{u\in\mathscr{U}_0}\Pr(U=u)\\&\qquad\times\Pr\left(\bigcup_{z\in Q_n}\mathscr{A}_z\cap\mathscr{B}_z\mid U=u\right)\\&\quad\leqslant\Pr(U\notin\mathscr{U}_0)+\sum_{u\in\mathscr{U}_0}\Pr(U=u)\\&\qquad\times\sum_{z\in Q_n}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u).\end{aligned} \quad (4.18)$$

Thus it is enough to prove the following three lemmas.

LEMMA 4.1. *For large enough n,*

$$\sum_{z\in Q_n}\Pr(\mathscr{A}_z\cap\bar{\mathscr{B}}_z)\leqslant e^{-\delta n}.$$

LEMMA 4.2. *For large enough n,*

$$\Pr(U\notin\mathscr{U}_0)\leqslant n^{-1+\delta}.$$

LEMMA 4.3. *For large enough n and $u\in\mathscr{U}_0$,*

$$\sum_{z\in Q_n}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u)\leqslant e^{-\delta n}.$$

4.3. Proofs of Lemmas 4.1 *and* 4.2

We first introduce the Central Limit Theorem for independent identically distributed (i.i.d.) unbiased Bernoulli random variables. (Random variables $Y_1, Y_2, \ldots$ are i.i.d. unbiased Bernoulli random variables if they are mutually independent, identically distributed and $\Pr(Y_1=1)=\Pr(Y_1=-1)=1/2$.) One might refer to any book on probability (e.g., [4]) for the proof. Also it is a good exercise to prove it directly from Stirling's Formula

$$n!=\sqrt{2\pi n}(n/e)^n\,e^{o_n},$$

where $1/(12n+1)<o_n<1/12n$.

For the rest of this section we refer to unbiased Bernoulli random variables simply as Bernoulli random variables.

THEOREM 4.1 (Central Limit Theorem for i.i.d. Bernoulli Random Variables). *Let $Y_1, Y_2, \ldots$ be i.i.d. Bernoulli random variables and*

$$S_m:=\sum_{i=1}^{m}Y_i \qquad \text{for} \quad m=1,2,\ldots.$$

Suppose $\phi\colon\mathbb{R}\to\mathbb{R}$ is a piecewise continuous function (i.e., continuous except possibly at a finite number of points) that satisfies

$$|\phi(t)|\leqslant ce^{(1-\rho)\,t^2/2}$$

for some positive constants ρ, c. Then we have

$$\mathrm{E}[\,\phi(m^{-1/2}S_m)\,]=(2\pi)^{-1/2}\int_{-\infty}^{\infty}\phi(t)\,e^{-t^2/2}\,dt+o(1),$$

where $o(1)$ goes to zero as m goes to infinity.

COROLLARY 4.2. *Let* $\{Y_{ij}\}_{i=1,\dots,r,\ j=1,\dots,m}$ *be a collection of i.i.d. Bernoulli random variables and let*

$$Y_i := m^{-1/2} \sum_{j=1}^{m} Y_{ij}, \qquad i = 1, \dots, r, \qquad Y := \sum_{i=1}^{r} |Y_i|.$$

Then, for sufficiently large r and m,

$$(1/r) \log \Pr(Y \geqslant \theta_0 r) \leqslant (1-\delta)\, h(\theta_0; \lambda_0)$$

and

$$(1/r) \log \Pr(Y \leqslant (1+\delta)(2/\pi)^{1/2} r) \leqslant -(1-\delta)\, v,$$

where $v := -\inf_{\lambda>0}\{\lambda^2/2 + f(\lambda) + \log 2 - (1+\delta)(2/\pi)^{1/2}\lambda\} > 0$.

Proof of Corollary 4.2. Theorem 4.1 for $\phi(t) = e^{-\lambda_0 |t|}$ implies

$$\begin{aligned}
\mathrm{E}[e^{-\lambda_0 |Y_i|}] &= \mathrm{E}[e^{-\lambda_0 |Y_1|}] \\
&= (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{-\lambda_0 |t|} e^{-t^2/2}\, dt + o(1) \\
&= \exp(\lambda_0^2/2 + \log(\Phi(-\lambda_0)) + \log 2 + o(1)) \\
&= \exp(\lambda_0^2/2 + f(-\lambda_0) + \log 2 + o(1)).
\end{aligned}$$

Since the Y_i are independent,

$$(1/r) \log \mathrm{E}[e^{-\lambda_0 Y}] = \lambda_0^2/2 + f(-\lambda_0) + \log 2 + o(1).$$

Thus the Markov inequality

$$\Pr(Y < y) \leqslant \mathrm{E}[e^{-\lambda(Y-y)}] \qquad \text{for all} \quad y \text{ and } \lambda \geqslant 0 \tag{4.19}$$

yields

$$\begin{aligned}
&(1/r) \log \Pr(Y \leqslant \theta_0 r) \\
&\quad \leqslant (1/r) \log \mathrm{E}[\exp(-\lambda_0 (Y - \theta_0 r))] \\
&\quad = \lambda_0^2/2 + f(-\lambda_0) + \log 2 + o(1) + \theta_0 \lambda_0 \\
&\quad = h(\theta_0; \lambda_0) + o(1) \leqslant (1-\delta)\, h(\theta_0; \lambda_0)
\end{aligned}$$

for sufficiently large r and m (note that $h(\theta_0; \lambda_0) < 0$).

Similarly, for all $\lambda > 0$,

$$\begin{aligned}
&(1/r) \log \Pr(Y \geqslant (1+\delta)(2/\pi)^{1/2} r) \\
&\quad \leqslant (1/r) \log \mathrm{E}[\exp(\lambda Y - \lambda(1+\delta)(2\pi)^{1/2} r)] \\
&\quad = \lambda^2/2 + f(\lambda) + \log 2 + o(1) - (1+\delta)(2/\pi)^{1/2} \lambda.
\end{aligned}$$

Thus

$$(1/r) \log \Pr(Y \geqslant (1+\delta)(2/\pi)^{1/2} r) \leqslant -(1-\delta)\, vr. \quad \blacksquare$$

Proof of Lemma 4.1. We prove Lemma 4.1 only for odd n. An analogous (but slightly more complicated) proof holds for even n.

Since the event $\mathscr{A}_z$ depends only on signatures $\sum z_j X_{ij}$ and $\mathscr{B}_z$ depends only on $|\sum z_j X_{ij}|$, $\mathscr{A}_z$ and $\mathscr{B}_z$ are independent for odd n and

$$\Pr(\mathscr{A}_z \cap \bar{\mathscr{B}}_z) = \Pr(\mathscr{A}_z)\, \Pr(\bar{\mathscr{B}}_z). \tag{4.20}$$

Thus

$$\Pr(\mathscr{A}_z) = 2^{-k} = 2^{-(1-\varepsilon)n}$$

gives

$$\Pr(\mathscr{A}_z \cap \bar{\mathscr{B}}_z) = 2^{-(1-\varepsilon)n} \Pr(\bar{\mathscr{B}}_z). \tag{4.21}$$

Because $\{z_j X_{ij}\}_{i=1,\dots,(1-\varepsilon)n,\ j=1,\dots,n}$ (for fixed $z \in Q_n$) are i.i.d. Bernoulli random variables, Corollary 4.2 (for $m = n$ and $r = (1-\varepsilon)n$) yields

$$\begin{aligned}
&\log \Pr(\bar{\mathscr{B}}_z) \leqslant (1-\delta)(1-\varepsilon)\, h(\theta_0; \lambda_0)\, n \\
&\qquad \text{for all} \quad z \in Q_n \text{ and sufficiently large } n.
\end{aligned} \tag{4.22}$$

Remark. The choice of sufficiently large n in (4.22) must be uniform in z because we will eventually consider $\sum_{z \in Q_n} \Pr(\bar{\mathscr{B}}_z)$. However, uniformity is guaranteed since all $\sum_{i=1}^{n} |\sum_{j=1}^{n} z_j X_{ij}|$ $(z \in Q_n)$ are identically distributed; in particular, all $\Pr(\bar{\mathscr{B}}_z)$ are the same.

Therefore, (4.14) and

$$\begin{aligned}
&\sum_{z \in Q_n} \Pr(\mathscr{A}_z \cap \bar{\mathscr{B}}_z) \\
&\qquad \leqslant 2^n 2^{-(1-\varepsilon)n} \exp((1-\delta)(1-\varepsilon)\, h(\theta_0; \lambda_0)\, n)
\end{aligned}$$

imply that

$$\begin{aligned}
&(1/n) \log \left(\sum_{z \in Q_n} \Pr(\mathscr{A}_z \cap \bar{\mathscr{B}}_z) \right) \\
&\qquad \leqslant \varepsilon \log 2 + (1-\delta)(1-\varepsilon)\, h(\theta_0; \lambda_0) < -\delta. \quad \blacksquare
\end{aligned}$$

Proof of Lemma 4.2. Since

$$\Pr(U \notin \mathscr{U}_1) = \Pr\left(\sum_{j=1}^{n} \left| \sum_{i=1}^{k} X_{ij} \right| > (1+\delta)(2k/\pi)^{1/2} n \right),$$

Corollary 4.2 (for $m = k$ and $r = n$) gives

$$(1/n) \log \Pr(U \notin \mathscr{U}_1) \leqslant -(1-\delta)\, v.$$

Thus it is enough to show

$$\Pr(U \notin \mathscr{U}_2) \leqslant n^{-1+\delta}/2.$$

Set

$$Z_j = \log(\cosh(\mu_0 k^{-1/2} U_j)), \qquad j = 1, 2, ..., n.$$

Then $\{Z_j\}_{j=1, ..., n}$ are i.i.d. and Theorem 4.1 yields

$$\begin{aligned} \mathrm{E}[Z_1] &= (2\pi)^{-1/2} \int_{-\infty}^{\infty} \log(\cosh(\mu_0 t))\, e^{-t^2/2}\, dt + o(1) \\ &= g(\mu_0) + o(1). \end{aligned}$$

Hence

$$\mathrm{E}\left[\sum_{j=1}^{n} Z_j\right] = n\,\mathrm{E}[Z_1] = (1 + o(1))\; g(\mu_0)\, n$$

and

$$\mathrm{Var}\left[\sum_{j=1}^{n} Z_j\right] = n\,\mathrm{Var}[Z_1].$$

Because $\mathrm{Var}[Z_1]$ tends to a finite number again by Theorem 4.1, Chebyshev's inequality implies that

$$\begin{aligned} \Pr(U \notin \mathscr{U}_2) &\leqslant \Pr\left(\sum_{j=1}^{n} Z_j - \mathrm{E}\left[\sum_{j=1}^{n} Z_j\right] > g(\mu_0)\, \delta n/2\right) \\ &\leqslant \frac{4\,\mathrm{Var}[\sum_{j=1}^{n} Z_j]}{(g(\mu_0)\, \delta n)^2} \\ &= \frac{4n\,\mathrm{Var}[Z_1]}{(g(\mu_0)\, \delta n)^2} = O(n^{-1}). \quad \blacksquare \end{aligned}$$

4.4. Proof of Lemma 4.3

The proof of Lemma 4.3 consists of three lemmas. The first is

LEMMA 4.4.

$$\Pr(\mathscr{A}_z \mid U = u) \leqslant \Pr(z \cdot X^{(1)} \geqslant 0 \mid U = u)^k$$
$$\textit{for all} \quad z \in Q_n \textit{ and } u \in \mathscr{U}.$$

For the proof of Lemma 4.4 we consider the n-dimensional integer lattice $\mathbb{L}_n$ for which

$$u \leqslant v \qquad \text{if and only if} \quad u_j \leqslant v_j, \qquad \forall j = 1, ..., n,$$

where u, v are n-dimensional integer vectors, and

$$u \wedge v = (\min\{u_j, v_j\})_{j=1, ..., n},$$
$$u \vee v = (\max\{u_j, v_j\})_{j=1, ..., n}.$$

Clearly, the lattice is distributive, and $\mathscr{U}$, Q_n are (finite) sublattices of $\mathbb{L}_n$.

Note that it is enough to show Lemma 4.4 for $z = \underline{1} := (1, ..., 1)$. Let $\mathscr{A}_l$, $l = 1, 2, ..., k$, be the (nested family of) events

$$\mathscr{A}_l := \{\underline{1} \cdot X^{(i)} \geqslant 0 \text{ for all } i = 1, 2, ..., l\},$$

and

$$U^{(l)} := (U_1 - X_{l1}, ..., U_n - X_{ln}).$$

The set of all possible (vector) values of $U^{(l)}$ is denoted by $\mathscr{U}^{(l)}$. Then $\mathscr{U}^{(l)}$ is also a sublattice of $\mathbb{L}_n$.

We claim that

$$u \leqslant v \Rightarrow \Pr(\mathscr{A}_l \mid U^{(l+1)} = u) \leqslant \Pr(\mathscr{A}_l \mid U^{(l+1)} = v)$$
$$\text{for all} \quad u, v \in \mathscr{U}^{(l+1)}, \qquad l = 1, ..., k-1. \tag{4.23}$$

This, together with the FKG inequality (see, e.g., [1, p. 74]), implies Lemma 4.4 as follows.

Proof of Lemma 4.4 *modulo* (4.23). Let

$$\Pr^{(u)}(\cdot) := \Pr(\cdot \mid U = u).$$

Since

$$\Pr^{(u)}(\underline{1} \cdot X^{(l+1)} \geqslant 0) = \Pr^{(u)}(\underline{1} \cdot X^{(1)} \geqslant 0)$$
$$\text{for all} \quad l = 1, ..., k-1,$$

the result follows if it is true that

$$\Pr^{(u)}(\mathscr{A}_{l+1}) \leqslant \Pr^{(u)}(\underline{1} \cdot X^{(l+1)} \geqslant 0)\, \Pr^{(u)}(\mathscr{A}_l)$$
$$\text{for all} \quad l = 1, ..., k-1. \tag{4.24}$$

Consider

$$\begin{aligned} &\Pr^{(u)}(\mathscr{A}_{l+1}) \\ &\quad = \sum_{x \in Q_n} \Pr^{(u)}(X^{(l+1)} = x)\, \Pr^{(u)}(\mathscr{A}_{l+1} \mid X^{(l+1)} = x) \\ &\quad = \sum_{x \in Q_n} \Pr^{(u)}(X^{(l+1)} = x)\, \chi(\underline{1} \cdot x \geqslant 0) \\ &\qquad \times \Pr^{(u)}(\mathscr{A}_l \mid X^{(l+1)} = x), \end{aligned} \tag{4.25}$$

where

$$\chi(\underline{1}\cdot x\geqslant 0):=\begin{cases}1 & \text{if } \underline{1}\cdot x\geqslant 0\\ 0 & \text{otherwise.}\end{cases}$$

Note that $\chi(\underline{1}\cdot x\geqslant 0)$ is an increasing function of x; that is,

$$x\leqslant x'\Rightarrow \chi(\underline{1}\cdot x\geqslant 0)\leqslant \chi(\underline{1}\cdot x'\geqslant 0).$$

Also, (4.23) implies that

$$\begin{aligned}\Pr^{(u)}(\mathscr{A}_l\mid X^{(l+1)}=x)&=\Pr(\mathscr{A}_l\mid U=u,\ X^{(l+1)}=x)\\&=\Pr(\mathscr{A}_l\mid U^{(l+1)}=u-x)\end{aligned}$$

is a decreasing function of x. Moreover,

$$\begin{aligned}\Pr^{(u)}(x)&:=\Pr^{(u)}(X^{(l+1)}=x)\\&=\prod_{j=1}^{n}\Pr(X_j^{(l+1)}=x_j\mid U_j=u_j)\end{aligned}$$

and

$$\Pr^{(u)}(x)\Pr^{(u)}(x')=\Pr^{(u)}(x\wedge x')\Pr^{(u)}(x\vee x')\quad\text{for all}\quad x,x'\in Q_n.$$

Thus the FKG inequality and (4.25) yield

$$\begin{aligned}&\Pr^{(u)}(\mathscr{A}_{l+1})\\&\leqslant\Big(\sum_{x\in Q_n}\Pr^{(u)}(X^{(l+1)}=x)\,\chi(\underline{1}\cdot x\geqslant 0)\Big)\\&\quad\times\Big(\sum_{x\in Q_n}\Pr^{(u)}(X^{(l+1)}=x)\Pr^{(u)}(\mathscr{A}_l\mid X^{(l+1)}=x)\Big)\\&=\Pr^{(u)}(\underline{1}\cdot X^{(l+1)}\geqslant 0)\Pr^{(u)}(\mathscr{A}_l).\quad\blacksquare\end{aligned}$$

Proof of (4.23). Note that it is enough to show (4.23) for v with $v_1=u_1+2$ and $v_j=u_j$ for $j\neq 1$.

Define the set M of all indices except $l+1$ such that $X_{i1}=1$,

$$M:=\{i\colon X_{i1}=1,\ 1\leqslant i\leqslant k,\ i\neq l+1\},$$

and

$$\begin{aligned}&\Pr(\mathscr{A}\mid M=I,\ U'=u')\\&\quad:=\Pr(\mathscr{A}\mid M=I,\ U_j^{(l+1)}=u_j\text{ for all }j=2,3,\ldots,n),\end{aligned}$$

where $U':=(U_2^{(l+1)},\ldots,U_n^{(l+1)})$. Then clearly

$$\begin{aligned}U_1^{(l+1)}=u_1\qquad&\text{if and only if the size}\\&|M|\text{ of }M\text{ is }(u_1+k-1)/2,\end{aligned}$$

and for $m:=(u_1+k-1)/2$ and $I\subseteq\{1,2,\ldots,l,l+2,\ldots,k\}$ with $|I|=m$,

$$\Pr(\mathscr{A}\mid M=I,\ U^{(l+1)}=u):=\Pr(\mathscr{A}\mid M=I,\ U'=u').$$

Since all I of size m are equally likely to be M, it is easy to see that

$$\begin{aligned}&\Pr(\mathscr{A}_l\mid U^{(l+1)}=u)\\&\quad=\binom{k-1}{m}^{-1}\sum_{I\colon|I|=m}\Pr(\mathscr{A}_l\mid M=I,\ U'=u')\end{aligned}\tag{4.26}$$

and similarly

$$\begin{aligned}&\Pr(\mathscr{A}_l\mid U^{(l+1)}=v)\\&\quad=\binom{k-1}{m+1}^{-1}\sum_{J\colon|J|=m+1}\Pr(\mathscr{A}\mid M=J,\ U'=u'),\end{aligned}\tag{4.27}$$

where all I, J are subsets of $\{1,\ldots,l,l+2,\ldots,k\}$.

Since

$$\begin{aligned}I\subset J\Rightarrow\Pr(\mathscr{A}_l\mid M=I,\ U'=u')\\\leqslant\Pr(\mathscr{A}_l\mid M=J,\ U'=u'),\end{aligned}\tag{4.28}$$

we have

$$\begin{aligned}&\sum_{\substack{I\colon|I|=m\\I\subset J}}\Pr(\mathscr{A}_l\mid M=I,\ U'=u')\\&\qquad\leqslant(m+1)\Pr(\mathscr{A}_l\mid M=J,\ U'=u'),\\&\qquad\text{for all }J\text{ with }|J|=m+1.\end{aligned}$$

Thus

$$\begin{aligned}&(k-1-m)\sum_{I\colon|I|=m}\Pr(\mathscr{A}_l\mid M=I,\ U'=u')\\&\quad=\sum_{I\colon|I|=m}\ \sum_{\substack{J\colon J\supset I\\|J|=m+1}}\Pr(\mathscr{A}_l\mid M=I,\ U'=u')\\&\quad=\sum_{J\colon|J|=m+1}\ \sum_{\substack{I\colon I\subset J\\|I|=m}}\Pr(\mathscr{A}_l\mid M=I,\ U'=u')\\&\quad\leqslant(m+1)\sum_{J\colon|J|=m+1}\Pr(\mathscr{A}\mid M=J,\ U'=u').\end{aligned}$$

Finally,

$$\begin{aligned}\binom{k-1}{m}^{-1} &\sum_{I:\,|I|=m} \Pr(\mathscr{A}_l \mid M=I,\, U'=u') \\ &\leqslant \frac{m+1}{k-1-m}\binom{k-1}{m}^{-1} \\ &\quad \times \sum_{J:\,|J|=m+1} \Pr(\mathscr{A} \mid M=J,\, U'=u') \\ &= \binom{k-1}{m+1}^{-1} \sum_{J:\,|J|=m+1} \Pr(\mathscr{A} \mid M=J,\, U'=u').\end{aligned}$$

This completes the proof by (4.26) and (4.27). ∎

The probability $\Pr(z \cdot X^{(1)} \geqslant 0 \mid U=u)$ in Lemma 4.4 is not hard to estimate since

$$\Pr(X_{1j}=1 \mid U=u) = \tfrac{1}{2}(1+u_j/k) \qquad \text{for all} \quad j=1, 2, ..., n, \tag{4.29}$$

and all events are mutually independent. Thus

$$\Pr(z_j X_{1j}=1 \mid U=u) = \tfrac{1}{2}(1+z_j u_j/k) \quad \text{for all } j=1, 2, ..., n, \tag{4.30}$$

and

$$\mathrm{E}[z \cdot X^{(1)} \mid U=u] = \sum_{j=1}^{n} z_j u_j/k = (z \cdot u)/k.$$

Hence it is reasonable to expect that the random variable

$$\frac{z \cdot X^{(1)} - (z \cdot u)/k}{\mathrm{Var}[z \cdot X^{(1)} \mid U=u]^{1/2}}$$

is asymptotically standard normal. Also, since we expect that

$$\mathrm{Var}[z \cdot X^{(1)} \mid U=u] = (1+o(1))\, n \qquad \text{for} \quad u \in \mathscr{U}_1,$$

we might easily have

$$\Pr(z \cdot X^{(1)} \geqslant 0 \mid U=u) = (1+o(1))\, \Phi\left(\frac{z \cdot u}{kn^{1/2}}\right). \tag{4.31}$$

However, we must have (4.31) uniformly in z and u as in the remark after (4.22). Thus extra effort is required.

Define, for a (fixed) constant θ,

$$W_n(\theta) := \Big\{ w=(w_1, ..., w_n):\ -1 \leqslant w_j \leqslant 1 \text{ for all } j,\ \sum_{j=1}^{n} w_j \leqslant \theta n^{1/2} \text{ and } \sum_{j=1}^{n} |w_j| \leqslant n^{3/4} \Big\}.$$

Let $w \in W_n(\theta)$ and let $\{ Y_j(w) \}_{j=1, 2, ..., n}$ be a family of mutually independent ± 1 random variables with

$$\Pr_0(Y_j(w)=1) = \tfrac{1}{2}(1+w_j),$$

and let

$$Y(w) := \sum_{j=1}^{n} Y_j(w).$$

The symbol $\Pr_0$ is used to distinguish the new probability space from our original probability space.

Since W_n is compact and the function

$$\begin{aligned}&\Pr_0(Y(w) \geqslant 0) \\ &\quad = 2^{-n} \sum_{l \geqslant n/2} \sum_{\substack{J:\,|J|=l \\ J \subset [n]}} \prod_{j \in J} (1+w_j) \prod_{j \notin J} (1-w_j)\end{aligned}$$

is continuous, there is a $w^{(n)} \in W_n(\theta)$ such that

$$p_n(\theta) := \sup_{w \in W_n(\theta)} \Pr_0(Y(w) \geqslant 0) = \Pr_0(Y(w^{(n)}) \geqslant 0).$$

Also it is routine to check that

$$\sum_{j=1}^{n} w_j^{(n)} = \theta n^{1/2}. \tag{4.32}$$

LEMMA 4.5.

$$\lim_{n \to \infty} p_n(\theta) = \Phi(\theta).$$

Proof. Let $w := w^{(n)}$ and define

$$Z_j := Y_j(w) - w_j,$$

and

$$\begin{aligned}S_n := Y(w) - \theta n^{1/2} &= \sum_{j=1}^{n} (Y_j(w) - w_j) \\ &= \sum_{j=1}^{n} Z_j \qquad (\text{by } (4.32)).\end{aligned} \tag{4.33}$$

Define the corresponding characteristic functions

$$\varphi_j(t) := \mathrm{E}_0[\exp(itZ_j)]$$

and

$$\psi_n(t) := \mathrm{E}_0[\exp(itn^{-1/2}S_n)],$$

where $i=\sqrt{-1}$. Because

$$\Pr_0(Y(w)\geqslant 0)=\Pr_0(n^{-1/2}S_n\geqslant -\theta),$$

it is enough to show that the distribution of $n^{-1/2}S_n$ converges (in distribution) to the standard normal distribution. Equivalently, by the Continuity Theorem (see, e.g., [4, p. 171]), we need to show that

$$\lim_{n\to\infty}\psi_n(t)=e^{-t^2/2}\quad\text{for all } t. \tag{4.34}$$

Note that (4.33) gives

$$\psi_n(t)=\prod_{j=1}^{n}\varphi_j(n^{-1/2}t)=\prod_{j=1}^{n}(1+(\varphi_j(n^{-1/2}t)-1)). \tag{4.35}$$

Since

$$|Z_j|\leqslant 2,\qquad \mathrm{E}_0[Z_j]=0,\qquad\text{and}\qquad \mathrm{E}_0[Z_j^2]=1-w_j^2,$$

Taylor's Theorem yields

$$\varphi_j(n^{-1/2}t)-1=-(1+o(1))\frac{t^2}{2n}(1-w_j^2). \tag{4.36}$$

Furthermore, (4.35), (4.36), and

$$|\log(1+s)-s|\leqslant 2s^2\qquad\text{for}\quad |s|\leqslant 1/2$$

yield

$$\begin{aligned}\log\psi_n(t)&=\sum_{j=1}^{n}\log(1+(\varphi_j(n^{-1/2}t)-1))\\&=-(1+o(1))\frac{t^2}{2n}\sum_{j=1}^{n}(1-w_j^2)+o(1).\end{aligned}$$

Finally,

$$\sum_{j=1}^{n}w_j^2\leqslant\sum_{j=1}^{n}|w_j|\leqslant n^{3/4}$$

implies that

$$\lim_{n\to\infty}\log\psi_n(t)=-\frac{t^2}{2}.\quad\blacksquare$$

COROLLARY 4.3. *For $\theta>0$ there is an integer $n(\theta)$ such that if $n\geqslant n(\theta)$ then*

$$\Pr(z\cdot X^{(1)}\geqslant 0\mid U=u)\leqslant(1+\delta)\,\Phi(\theta)$$
$$\textit{for all}\quad z\in Q_n,\qquad u\in\mathscr{U}_1\quad\text{with}\quad z\cdot u\leqslant\theta n^{1/2}k.$$

Proof. From $u\in\mathscr{U}_1$ and (4.29), it is easy to see that

$$\Pr(z\cdot X^{(1)}\geqslant 0\mid U=u)\leqslant p_n(\theta).$$

The result follows from Lemma 4.5. ∎

Let

$$\sum_1^{(u)}:=\sum_{\substack{z:\,z\in Q_n\\ z\cdot u<\theta_0 n^{1/2}k}}\qquad\text{and}\qquad\sum_2^{(u)}:=\sum_{\substack{z:\,z\in Q_n\\ z\cdot u\geqslant\theta_0 n^{1/2}k}}.$$

Then

$$\begin{aligned}\sum_{z\in Q_n}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u)&=\sum_1^{(u)}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u)\\&\quad+\sum_2^{(u)}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u).\end{aligned} \tag{4.37}$$

If $\mathscr{A}_z\cap\mathscr{B}_z$ occurs then (4.3) yields

$$z\cdot U=\sum_{i=1}^{n}\left|\sum_{j=1}^{n}z_jX_{ij}\right|\geqslant\theta_0 n^{1/2}k.$$

This, together with (4.37) and Lemma 4.4, implies that

$$\begin{aligned}\sum_{z\in Q_n}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u)&=\sum_2^{(u)}\Pr(\mathscr{A}_z\cap\mathscr{B}_z\mid U=u)\\&\leqslant\sum_2^{(u)}\Pr(\mathscr{A}_z\mid U=u)\\&\leqslant\sum_2^{(u)}\Pr(z\cdot X^{(1)}\mid U=u)^k.\end{aligned} \tag{4.38}$$

Let $\theta_l:=\theta_0+l\delta$ and

$$Q_l(u):=\{z\in Q_n:\,\theta_{l-1}n^{1/2}k\leqslant z\cdot u<\theta_l n^{1/2}k\},$$
$$l=1,2,\ldots,m_\delta,$$

where

$$m_\delta:=1+\left\lfloor\frac{(1+\delta)(1-\varepsilon)^{-1/2}(2/\pi)^{1/2}-\theta_0}{\delta}\right\rfloor.$$

Then (4.38) and Corollary 4.3 imply that, for all $u\in\mathscr{U}_1$ and $n>\max\{n(\theta_l):\,l=1,2,\ldots,m_\delta\}$, we have

$$\sum_{z \in Q_n} \Pr(\mathscr{A}_z \cap \mathscr{B}_z \mid U = u)$$
$$\leqslant \sum_{l=1}^{m_\delta} \sum_{z \in Q_l(u)} ((1+\delta)\, \Phi(\theta_l))^k$$
$$\leqslant m_\delta (1+\delta)^k \max\{|Q_l(u)|(\Phi(\theta_l))^k : l = 1, 2, ..., m_\delta\}. \tag{4.39}$$

Our final lemma is

LEMMA 4.6. *For all $u \in \mathscr{U}_2$ and $l = 1, 2, ..., m_\delta$,*

$$(1/n) \log |Q_l(u)|$$
$$\leqslant -(1-\varepsilon)^{1/2}\, \theta_{l-1} \mu_0 + (1+\delta)\, g(\mu_0) + \log 2.$$

Lemma 4.3 easily follows from Lemma 4.6 because (4.39) (recall $k = (1-\varepsilon)\, n$), (4.16), and Lemma 4.6 give

$$(1/n) \log \left(\sum_{z \in Q_n} \Pr(\mathscr{A}_z \cap \mathscr{B}_z \mid U = u) \right)$$
$$\leqslant (1/n) \log m_\delta + (1-\varepsilon) \log(1+\delta)$$
$$+ \max_{l \in \{1, ..., m_\delta\}} \{(1/n) \log |Q_l(u)| + (1-\varepsilon) \log(\Phi(\theta_l))\}$$
$$\leqslant \delta + \max_{l \in \{1, ..., m_\delta\}} \{-(1-\varepsilon)^{1/2}\, \theta_{l-1} \mu_0 + (1+\delta)\, g(\mu_0)$$
$$+ \log 2 + (1-\varepsilon)\, f(\theta_{l-1} + \delta)\}$$
$$< -\delta$$

for sufficiently large n.

Proof of Lemma 4.6. To prove Lemma 4.6, let $\{Z_j\}_{j=1, ..., n}$ be i.i.d. Bernoulli random variables. We use $\Pr^*(\cdot)$ for this new probability space. It is easy to see that

$$|Q_l(u)| = 2^n \Pr^* \left(\theta_{l-1} n^{1/2} k \leqslant \sum_{j=1}^n u_j Z_j < \theta_l n^{1/2} k \right)$$
$$\leqslant 2^n \Pr^* \left(\sum_{j=1}^n k^{-1/2} u_j Z_j \geqslant \theta_{l-1} (nk)^{1/2} \right)$$
$$= 2^n \Pr^* \left(\sum_{j=1}^n k^{-1/2} u_j Z_j \geqslant (1-\varepsilon)^{1/2}\, \theta_{l-1} n \right).$$

Define

$$\theta := (1-\varepsilon)^{1/2}\, \theta_{l-1} \qquad \text{and} \qquad v_j := k^{-1/2} u_j.$$

Then

$$(1/n) \log |Q_l(u)| \leqslant \log 2 + (1/n) \log \Pr^* \left(\sum_{j=1}^n v_j Z_j \geqslant \theta n \right). \tag{4.40}$$

Suppose $u \in \mathscr{U}_2$. Since

$$\mathrm{E}^* \left[\exp\left(\lambda \sum_{j=1}^n v_j Z_j \right) \right] = \prod_{j=1}^n \frac{e^{\lambda v_j} + e^{-\lambda v_j}}{2}$$
$$= \prod_{j=1}^n \cosh(\lambda v_j) \qquad \text{for all} \quad \lambda \in \mathbb{R},$$

it is easily seen that

$$(1/n) \log \Pr^* \left(\sum_{j=1}^n v_j Z_j \geqslant \theta n \right)$$
$$\leqslant (1/n) \log \mathrm{E}^* \left[\exp\left(\mu_0 \left(\sum_{j=1}^n v_j Z_j - \theta n \right) \right) \right]$$
$$\leqslant -\theta \mu_0 + (1/n) \log \mathrm{E}^* \left[\exp\left(\mu_0 \sum_{j=1}^n v_j Z_j \right) \right]$$
$$= -\theta \mu_0 + \frac{1}{n} \sum_{j=1}^n \log(\cosh(\mu_0 v_j))$$
$$\leqslant -\theta \mu_0 + (1+\delta)\, g(\mu_0).$$

Therefore, (4.40) yields

$$(1/n) \log |Q_l(u)| \leqslant -\theta \mu_0 + (1+\delta)\, g(\mu_0) + \log 2. \quad \blacksquare$$

5. LOWER BOUND: PROOF OF THEOREM 3.2

5.1. *Construction Used to Prove Theorem* 3.2

In this subsection we describe the construction used to prove Theorem 3.2. The next subsection will have a plausibility argument for the theorem, and then the theorem will be proved rigorously in the succeeding subsections.

If we were willing to settle for a sublinear lower bound on the critical value of k as a function of n, we could simply choose each component z_j of z to be the majority element (-1 or $+1$) from the corresponding column $(X_j^{(1)}, ..., X_j^{(k)})^T$ of the matrix X. Using this construction of z, it is not hard to show that for an appropriate constant $C_0 > 0$, if $0 < C < C_0$ and $k = \lfloor Cn/\log n \rfloor$,

$$\Pr(\exists z \in Q_n \text{ s.t. } X^{(i)} \cdot z > 0 \ \forall i = 1, 2, ..., k) \to 1$$

as $n \to \infty$. For a derivation of the best possible value of C_0, see [35]. For related results, see also [6–8].

We obtain the stronger (linear) lower bound of Theorem 3.2 by using a multi-stage procedure to construct a vector z such that $Xz > 0$ with high probability (w.h.p.). The construction takes place in $\Theta(\log \log n)$ steps as below. The basic idea in each step, or stage, is to examine a block of adjacent columns of X that have not yet been observed and to choose the corresponding components of z so as to reduce the number of rows of X having small cumulative inner product with z.

W.h.p., we ultimately obtain a vector z of full dimension (n) such that *no* row of X has small inner product with z. Our procedure is similar in spirit to that used by Komlós in his original proof that the probability of a random $n \times n$ binary matrix being nonsingular over the reals approaches 1 as $n \to \infty$. (See [3].)

Remark. For now we use the phrase "w.h.p." and the symbols $\approx$ and $\lesssim$ without defining them precisely. Roughly, one may take "w.h.p." to mean "w.p. $1 - O(n^{-\alpha})$ for every $\alpha > 0$." The symbol $\approx$ can mean either "is approximately equal to" or "has approximate distribution." Unless indicated otherwise, logarithms are base e.

Before describing how to choose z, let us define the quantities N, $\{k_j\}_{0 \leqslant j \leqslant N}$, $\{n_j\}_{0 \leqslant j \leqslant N}$, and introduce some additional notation. N will be the total number of stages (excluding the first) in the construction of z, and the stages will be numbered 0, 1, ..., N. In each stage j, for $0 \leqslant j \leqslant N$, we will focus on k_j of the k rows, using these rows to construct the next n_j entries of z, where $\sum_{j=0}^{N} n_j = n$.

DEFINITION. For all sufficiently large n, let

$$N := \max\{m \in Z: 10^{2^m} \leqslant n^{0.01}\} = \lfloor \log_2(0.01 \log_{10} n) \rfloor \quad (\sim \log\log n / \log 2). \tag{5.41}$$

(We suppress the dependence of N on n.)

DEFINITION. We define the fractions

$$f_0 = 1, \qquad f_1 = 1/200, \qquad f_j = 10^{-2^j} \quad \text{for} \quad 2 \leqslant j \leqslant N, \tag{5.42}$$

and let

$$A = \sum_{j=0}^{N} f_j. \tag{5.43}$$

Now let

$$n_0 = \lfloor n/A \rfloor, \qquad n_j = \left\lfloor (n/A) \sum_{i=0}^{j} f_i \right\rfloor - \left\lfloor (n/A) \sum_{i=0}^{j-1} f_i \right\rfloor \quad \text{for} \quad 1 \leqslant j \leqslant N. \tag{5.44}$$

Note. We see that $n_j \approx f_j n$ for all $j \in \{0, 1, ..., N\}$, and that

$$\sum_{j=0}^{N} n_j = \lfloor (n/A)\, A \rfloor = n.$$

Also note that $A \approx 1$, so that $n_0 \approx n$.

DEFINITION. For $0 \leqslant i \leqslant j \leqslant N$, let $X(i:j)$ be the submatrix of X formed by keeping only columns $(\sum_{r=0}^{i-1} n_r + 1)$ through $(\sum_{r=0}^{j} n_r)$ of X, i.e., the columns considered in Step i through Step j. Recalling that $X_m^{(\ell)} = X_{\ell m}$ for $1 \leqslant \ell \leqslant k$, $1 \leqslant m \leqslant n$, we let $X^{(r)}(i:j)$ be the rth row of $X(i:j)$.

DEFINITION. For $0 \leqslant i \leqslant j \leqslant N$, let $z(i:j)$ be the subvector formed by components $(\sum_{r=0}^{i-1} n_r + 1)$ through $(\sum_{r=0}^{j} n_r)$ of the vector z, where z will be constructed such that $Xz > 0$ w.h.p.

DEFINITION. For $1 \leqslant r \leqslant k$, $0 \leqslant i \leqslant j \leqslant N$, let

$$S^{(r)}(i:j) := X^{(r)}(i:j) \cdot z(i:j) \tag{5.45}$$

be the contribution to the inner product $X^{(r)} \cdot z$ from the columns considered in Step i through Step j.

For $0 \leqslant s \leqslant N$, during Step s we will examine $X(s:s)$ and use this submatrix in constructing the subvector $z(s:s)$. At the end of Step N, we will have examined the entire matrix $X(0:N) = X$ and constructed the entire vector $z(0:N) = z$.

DEFINITION. For all sufficiently large n, for $0 \leqslant s \leqslant N$, define k_s (the number of rows examined in Step s) as

$$\begin{aligned} k_0 &= k = 2\lfloor (1/2)(n/200) \rfloor + 1, \\ k_1 &= 2\lfloor (1/2)(n/10^8) \rfloor + 1, \\ k_s &= 2\lfloor (1/2)(f_s^3 n) \rfloor + 1 \quad \text{for} \quad 2 \leqslant s \leqslant N. \end{aligned} \tag{5.46}$$

Thus each k_s is odd, with $k_0 \approx n/200$, $k_1 \approx n/10^8$, and $k_s \approx n/1000^{2^s}$ for $s \geqslant 2$.

With all of these definitions out of the way, we can now give the algorithm for choosing the vector z. In Step 0, choose $z(0:0)$ as follows: For $1 \leqslant j \leqslant n_0$, let $z_j \in \{-1, 1\}$ have the same sign as the jth column sum of X; i.e.,

$$z_j = \operatorname{sgn}\left(\sum_{i=1}^{k} X_{ij} \right). \tag{5.47}$$

Equivalently, select each component z_j by taking a majority vote within the jth column of the matrix X.

For Step s, where $1 \leqslant s \leqslant N$, look at the n_s columns of $X(s:s)$, and restrict attention to the k_s rows yielding the smallest values for the (partial) inner products $S^{(r)}(0:s-1)$. (If there is any ambiguity in choosing the k_s rows, resolve the ambiguity arbitrarily.) Each component of $z(s:s)$ is chosen by taking a majority vote in the corresponding columns of $X(s:s)$, giving suffrage only to the k_s rows described above. (Since each k_s is odd, each column will have a strict majority.) Intuitively, this procedure will boost small partial inner products while adding unbiased noise to the others. (If k_s were allowed to be even, essentially the same results would be obtained by flipping a fair coin to break ties.)

In order to refer more easily to the row and column subsets of interest, we make the following definition.

DEFINITION. For $0 \leqslant s \leqslant N$, let

$$\mathscr{I}_s := \{r\text{: Row } r \text{ enters into the majority vote in Step } s\}. \tag{5.48}$$

Similarly, let

$$\mathscr{J}_s := \left\{ j\text{: } \sum_{i=0}^{s-1} n_i + 1 \leqslant j \leqslant \sum_{i=0}^{s} n_i \right\} \tag{5.49}$$

be the set of column indices involved in Step s.

Remark. Note that

$$|\mathscr{I}_s| = k_s \qquad \text{and} \qquad |\mathscr{J}_s| = n_s. \tag{5.50}$$

Also,

$$z_j = \operatorname{sgn}\left(\sum_{i \in \mathscr{I}_s} X_{ij} \right) \qquad \text{for all} \quad j \in \mathscr{J}_s. \tag{5.51}$$

We take k_s to be odd for all s so that $z_j \in \{-1, 1\}$. If k is even, so that k_0 would be even, we can simply add an auxiliary row with independent and random entries to make k odd. To analyze the algorithm for choosing z, we will consider a sequence of decreasing positive threshold T_0, T_1, ..., T_N, and argue that w.h.p. the number of rows r of $X(0:s)$ with cumulative inner product $S^{(r)}(0:s) < T_s$, decreases rapidly with each step s. Finally, after Step N, we shall see that w.h.p. *no* row of $X(0:N)$ has $S^{(r)}(0:N) < T_N$, and hence all rows of X have positive inner product with z. We choose the thresholds T_j as below.

DEFINITION. For $0 \leqslant j \leqslant N$, let

$$T_j := \sqrt{n}/2^{j-1}. \tag{5.52}$$

It will also be convenient to define.

$$T_{-1} := +\infty.$$

Remark. For simplicity, we shall sometimes represent the set $\{1, 2, ..., m\}$ as $[m]$.

5.2. *Overview of Proof of Theorem* 3.2

In this overview of the proof of Theorem 3.2, we freely make use of the fact that under appropriate conditions, a large collection of identically distributed random variables has (w.h.p.) empirical distribution, or sample distribution, "close" to the common statistical distribution of its members. We will initially gloss over the difficulties caused by dependence of random variable (r.v.'s).

Since we wish to argue eventually that w.h.p. $S^{(r)}(0:N) > T_N > 0$, $\forall r \in [k]$, we shall compute (approximate) upper bounds on quantities such as $\Pr(S^{(r)}(i:j) \leqslant \eta)$ for various values of η. Such upper bounds correspond to *lower* bounds on the "typical" values of $S^{(r)}(i:j)$.

Throughout this section we use the term "with high probability" (or "w.h.p.") loosely. We may think of it as meaning "with probability $1 - O(n^{-\alpha})$ for every $\alpha > 0$." Letting t be the number of rows or columns that we consider together during a particular stage or substage of the proof, we generally have t comparable to k, n, k_s, or n_s, all of which lie between $n^{0.97}$ and n for every stage s. This fact will enable us to get uniform bounds on the rates of convergence of various "error" probabilities to 0 as $n \to \infty$, after which a union bound will use the fact that the *sum* of all such probabilities converges sufficiently quickly to 0 as $n \to \infty$.

Theorem 3.2 will be proved in detail in later sections. However, the primary idea behind the proof involves keeping track of the cumulative (partial) inner products $S^{(r)}(0:s)$. It will be useful to consider the empirical distribution of $S^{(r)}(i:j)$ for $r \in [k]$ for given values i, j with $0 \leqslant i \leqslant j \leqslant N$, especially when $i = 0$ or j.

DEFINITION. Given any nonempty set of row indices $\mathscr{A} \subseteq \{1, ..., k\}$, let the r.v. R (defined on the probability space Ω_2 with probability measure $\Pr_2(\cdot)$) be chosen uniformly from the elements of $\mathscr{A}$, and let

$$S(i:j \mid \mathscr{A}) := S^{(R)}(i:j).$$

Also let

$$S(i:j) := S(i:j \mid \{1, ..., k\}).$$

Remarks. For each nonempty $\mathscr{A} \subseteq [k]$, the quantity $S(i:j \mid \mathscr{A})$ depends both on the value of ω from the original probability space Ω (on which the X_{ij} are defined) and on the value of ω_2 from the newly defined probability space Ω_2.

For fixed $\omega \in \Omega$, $S(i:j \mid \mathscr{A})$ has statistical distribution (over Ω_2) equal to the *sample* distribution of the $|\mathscr{A}|$ integers $\{S^{(r)}(i:j)\}_{r \in \mathscr{A}}$. Thus for fixed $\omega \in \Omega$, the function $\Pr_2(S(i:j \mid \mathscr{A}) \leqslant u)$ of the real variable u is the *empirical*, or *sample*, cumulative distribution function (c.d.f.) of $S(i:j \mid \mathscr{A})$ and is equal to $|\{r \in \mathscr{A}\text{: } S^{(r)}(i:j) \leqslant u\}|/|\mathscr{A}|$.

For each fixed real value of u, the quantity $\Pr_2(S(i:j|\mathscr{A}) \leqslant u)$ is an ordinary r.v. over Ω, the original probability space.

When we use the term "w.h.p." without specifying a probability space, it will always be with respect to the original probability measure $\Pr(\cdot)$ over Ω.

Note that for any set $\mathscr{A}$ and its complement $\bar{\mathscr{A}} := \{1, ..., k\} \setminus \mathscr{A}$, the r.v. $S(s_1 : s_2)$ (defined on $\Omega \times \Omega_2$) may be

written as a mixture of the r.v.'s $S(s_1 : s_2 \mid \mathscr{A})$ and $S(s_1 : s_2 \mid \bar{\mathscr{A}})$, with weights $|\mathscr{A}|/k$ and $(k - |\mathscr{A}|)/k$, respectively. We will be particularly interested in the quantities

$$S(0 : j \mid \mathscr{A}),\ S(j+1 : j+1 \mid \mathscr{A}) \quad \text{and} \quad S(0 : j+1 \mid \mathscr{A}),$$

where $\mathscr{A} = [k]$, $\mathscr{I}_{j+1}$, or $\bar{\mathscr{I}}_{j+1}$.

We shall see by an inductive argument on s that, for all sufficiently large n, after each Step s, the empirical c.d.f. of $\{S^{(r)}(0 : s)\}_{r \in [k]}$ may be probabilistically upper-bounded as follows: w.h.p.,

$$\Pr_2(S(0 : s) \leqslant \eta) \lesssim \Phi\left(\frac{\eta - \mu_{s,1}}{\sigma_{s,1}}\right) + (1+\epsilon)\,\Phi\left(\frac{\eta - \mu_{s,2}}{\sigma_{s,2}}\right) \tag{5.53}$$

for all $\eta \leqslant T_s$, $s \in \{0, 1, ..., N\}$, where

$$\Phi(\lambda) := (2\pi)^{-1/2} \int_{-\infty}^{\lambda} e^{-t^2/2}\, dt,$$

$$\mu_{s,1} \approx T_{s-1}, \qquad \sigma_{s,1}^2 \approx n_s, \qquad \mu_{s,2} \approx \sqrt{2/(\pi k_s)}\, n_s,$$

$$\sigma_{s,2}^2 \approx \sum_{j=0}^{s} n_j \approx n \qquad \text{and} \qquad 0 \leqslant \epsilon < 1.$$

Remark. For $f(s, \eta)$ and $g(s, \eta)$ nondecreasing and nonnegative functions of η, the expression "$f(s, \eta) \lesssim g(s, \eta)$ for all $\eta \leqslant T_s$" may be taken roughly to mean that

$$f(s, \eta) \leqslant \delta_s + g(s, \eta) \qquad \text{for all} \quad \eta \leqslant T_s,$$

where $\delta_s \ll g(s, T_s)$.

Setting $\eta = T_s$ in (5.53), we shall see that, w.h.p.,

$$\Pr_2(S(0 : s) \leqslant T) := (1/k)\, |\{r : S^{(r)}(0 : s) \leqslant T_s\}|$$
$$\leqslant k_{s+1}/k \qquad \text{for all} \quad s \in \{0, 1, ..., N-1\}$$

and

$$|\{r : S^{(r)}(0 : N) \leqslant 0\}| = 0.$$

(The fact that (w.h.p.) $\Pr_2(S(0 : s) \leqslant T_s) \leqslant k_{s+1}/k$ will be critical.) Thus when the entire vector $z := z(0 : N)$ has been built up, we shall have $Xz > 0$ w.h.p.

We begin our heuristic derivation of (5.53) by considering the basis of the induction, with $s = 0$. Taking $\mu_{0,1} = T_{-1} = +\infty$, we have

$$\Phi\left(\frac{\eta - \mu_{s,1}}{\sigma_{s,1}}\right) = 0 \qquad \text{for all} \quad \eta \in \mathbb{R} \qquad \text{when} \quad s = 0,$$

and thus for $s = 0$ we need only show that, w.h.p.,

$$\Pr_2(S(0 : 0) \leqslant \eta) \lesssim (1+\epsilon)\,\Phi\left(\frac{\eta - \sqrt{2/(\pi k_0)}\, n_0}{\sqrt{n_0}}\right)$$
$$\text{for all} \quad \eta \leqslant T_0 := 2\sqrt{n}. \tag{5.54}$$

But for any given row $r \in \{1, ..., k\}$, $S^{(r)}(0 : 0)$ is the sum of n_0 i.i.d. Bernoulli r.v.'s $z_j X_j^{(r)}$ for $1 \leqslant j \leqslant n_0$, where $z_j = \operatorname{sgn}(\sum_{r=1}^{k} X_j^{(r)})$. For all sufficiently large odd k_0 (equivalently, for all sufficiently large n), it is easy to check that for each r, $\Pr(z_j = X_j^{(r)}) \approx 1/2 + 1/\sqrt{2\pi k_0}$.

DEFINITION. For $r \in [k]$, $j \in [n]$, let

$$w_j^{(r)} := w_{rj} := z_j X_j^{(r)} := z_j X_{rj}. \tag{5.55}$$

Using the definition of w_{rj} and the fact that each k_s is odd, we see that for all $r \in [k]$, $j \in [n]$,

$$w_{rj} \approx \text{Bern}(1/2 + \gamma_0/2) \qquad \text{with} \quad \gamma_0 \approx \sqrt{2/(\pi k_0)}. \tag{5.56}$$

By this we mean that each r.v. w_{rj} is ± 1 and approximately Bernoulli distributed with the given parameter, i.e.,

$$w_{rj} = \begin{cases} +1 & \text{w.p.} \quad \approx 1/2 + \gamma_0/2 \\ -1 & \text{w.p.} \quad \approx 1/2 - \gamma_0/2 \end{cases}. \tag{5.57}$$

For a more precise estimate of w_{rj}, see [8].

The following definition will be useful later when we prove Theorem 3.2 rigorously.

DEFINITION. For $0 \leqslant s \leqslant N$, let

$$\gamma_s := (1 - k_s^{-1/8})\sqrt{2/(\pi k_s)}. \tag{5.58}$$

Returning to the basis of the induction, we see that for each row r, with $1 \leqslant r \leqslant k$, $S^{(r)}(0 : 0)$ is the sum of n_0 i.i.d. Bernoulli r.v.'s each with mean $\mu \approx \gamma_0$ and variance $\sigma^2 \approx 1$ for large n. By a suitable version of the central limit theorem (CLT), it follows that for each $r \in [k]$,

$$S^{(r)}(0 : 0) \approx N(\gamma_0 n_0, n_0). \tag{5.59}$$

By this we mean that each r.v. $S^{(r)}(0 : 0)$ has a c.d.f. close to that of a normal distribution with mean $\gamma_0 n_0$ and variance (*not* standard deviation) $n_0 = (\sqrt{n_0})^2$. If the r.v.'s $\{S^{(r)}(0 : 0)\}_{r \in [k]}$ were all independent, it would follow by a suitable law of large numbers (LLN) that, w.h.p., the *empirical* distribution of $\{S^{(r)}(0 : 0)\}_{r \in [k]}$ is close to the common statistical distribution and thus satisfies

$$\Pr_2(S(0 : 0) \leqslant \eta) \lesssim \Phi((\eta - \gamma_0 n_0)/\sqrt{n_0}), \qquad \forall \eta \in \mathbb{R},$$

as we wished to show.

In reality, the r.v.'s $\{S^{(r)}(0:0)\}_{r\in[k]}$ are not quite independent, because for each $j\in\mathscr{J}_0(=\{1,...,n_0\})$, the r.v.'s w_{rj} are weakly dependent. This problem will be overcome by considering approximately $k_0^{1/10}=k^{1/10}$ rows at a time. Given any subset $\mathscr{B}\subseteq[k]$ with $|\mathscr{B}|\leqslant k^{1/10}$, we can define a set of r.v.'s $\{y_{rj}\}_{r\in\mathscr{B},\, j\in\mathscr{J}_0}$ on the same probability space as the w_{rj}'s, where $y_{rj}\leqslant w_{rj}$ for all r, j, and where the $k_0 n_0$ r.v.'s y_{rj} are mutually i.i.d. with distribution Bern $((1+\gamma_0)/2)$. The basis of the induction for (5.53) then follows readily. Finally, we check that (w.h.p.) $\Pr_2(S(0:s)\leqslant T_s)\lesssim k_{s+1}/k$ for $s=0$, as required to make the induction work. Setting

$$\eta=T_0:=2\sqrt{n},\qquad n_0\approx n,\qquad \text{and}$$
$$\gamma_0\approx\sqrt{2/(\pi k_0)}\approx 0.80/\sqrt{n/200}, \tag{5.60}$$

it is straightforward to show that (w.h.p.) for $s=0$,

$$\Pr_2(S(0:s)\leqslant T_s)\lesssim\Phi((2\sqrt{n}-0.80\sqrt{200}\sqrt{n})/\sqrt{n})$$
$$\approx\Phi(-9.3)\leqslant k_{s+1}/k, \tag{5.61}$$

where the last inequality follows because $\Phi(-9.3)<(1/\sqrt{2\pi}(9.3))\exp(-(9.3)^2/2)<10^{-20}$ and $k_1/k\approx 2\times 10^{-6}$ from (5.46).

Now we sketch the proof of the induction step. We suppose that (5.53) holds w.h.p. for $0\leqslant s\leqslant j$, where $j\in\{0,...,N-1\}$, and use this supposition to show that (5.53) also holds w.h.p. for $s=j+1$. By hypothesis, w.h.p.,

$$\Pr_2(S(0:j)\leqslant\eta)$$
$$\lesssim\Phi((\eta-\mu_{j,1})/\sigma_{j,1})+(1+\epsilon)\,\Phi((\eta-\mu_{j,2})/\sigma_{j,2}) \tag{5.62}$$

for all $\eta\leqslant T_j:=\sqrt{n}/2^{j-1}$, where

$$\mu_{j,1}\approx T_{j-1}=\sqrt{n}/2^{j-2},\qquad \sigma_{j,1}^2\approx n_j,$$
$$\mu_{j,2}\approx\sqrt{2/(\pi k_j)}\,n_j,\qquad\text{and}\qquad \sigma_{j,2}^2\approx n.$$

In establishing the induction step for any given j, there are two cases to consider, depending on whether $S^{(r)}(0:j)$ is among the smallest k_{j+1} accumulated inner products (in which case $r\in\mathscr{I}_{j+1}$) or not (in which case $r\in\bar{\mathscr{I}}_{j+1}:=\{1,...,k\}\setminus\mathscr{I}_{j+1}$).

It follows from arguments similar to those used to establish the basis of the induction that

$$S^{(r)}(j+1:j+1)\approx N(\gamma_{j+1}n_{j+1},n_{j+1})\qquad\text{for all}\quad r\in\mathscr{I}_{j+1} \tag{5.63}$$

and

$$S^{(r)}(j+1:j+1)\approx N(0,n_{j+1})\qquad\text{for all}\quad r\in\bar{\mathscr{I}}_{j+1}. \tag{5.64}$$

In addition, the r.v.'s above are approximately independent for all sufficiently large n. Furthermore, it is easily seen that for every $r_1, r_2\in\{1,...,k\}$ (with r_1 and r_2 possibly equal), if we condition upon whether r_2 is in $\mathscr{I}_{j+1}$ or $\bar{\mathscr{I}}_{j+1}$, then $S^{(r_1)}(0:j)$ and $S^{(r_2)}(j+1:j+1)$ are conditionally independent. As before we can show that w.h.p. all sample distributions of interest are close to their statistical distributions. Thus w.h.p.

$$S(j+1:j+1\mid\mathscr{I}_{j+1})\approx N(\gamma_{j+1}n_{j+1},n_{j+1}) \tag{5.65}$$

and

$$S(j+1:j+1\mid\bar{\mathscr{I}}_{j+1})\approx N(0,n_{j+1}), \tag{5.66}$$

where

$$\gamma_{j+1}\approx\sqrt{2/(\pi k_{j+1})}$$

and the distributions are defined on the probability space Ω_2.

Now we combine the facts above with the induction hypothesis (including the assertion that, w.h.p., $\Pr_2(S(0:j)\leqslant T_j)\leqslant k_{j+1}/k$) to complete the induction. The inequality just mentioned implies that, w.h.p.,

$$S^{(r)}(0:j)>T_j,\qquad\forall r\in\bar{\mathscr{I}}_{j+1}. \tag{5.67}$$

(In words, w.h.p. all of the rows except the k_{j+1} rows with smallest partial inner product after Step j have partial inner product greater than the threshold T_j.) Recalling that $|\mathscr{I}_{j+1}|=k_{j+1}$, $|\bar{\mathscr{I}}_{j+1}|=k-k_{j+1}$, we have the following identity involving empirical c.d.f.'s:

$$\Pr_2(S(0:j+1)\leqslant\eta)$$
$$=\frac{k_{j+1}}{k}\Pr_2(S(0:j+1\mid\mathscr{I}_{j+1})\leqslant\eta)$$
$$+\left(1-\frac{k_{j+1}}{k}\right)\Pr_2(S(0:j+1\mid\bar{\mathscr{I}}_{j+1})\leqslant\eta). \tag{5.68}$$

Using the conditional independence of $S^{(r_1)}(0:j)$ and $S^{(r_2)}(j+1:j+1)$—conditioned on whether r_2 is in $\mathscr{I}_{j+1}$ or $\bar{\mathscr{I}}_{j+1}$—together with some limit theorems relating statistical and sample distributions, we can bound the second term as follows: w.h.p., for all $\eta\leqslant T_j$,

$$\left(1-\frac{k_{j+1}}{k}\right)\Pr_2(S(0:j+1\mid\bar{\mathscr{I}}_{j+1})\leqslant\eta)$$
$$\lesssim 1\cdot\Pr_2(S(0:j\mid\bar{\mathscr{I}}_{j+1})+S(j+1:j+1\mid\bar{\mathscr{I}}_{j+1})\leqslant\eta)$$
$$\leqslant\Pr_2(T_j+S(j+1:j+1\mid\bar{\mathscr{I}}_{j+1})\leqslant\eta)$$
$$\lesssim\Phi((\eta-T_j)/\sqrt{n_{j+1}}).$$

The first term on the right in (5.68) can be (approximately) bounded above as follows: w.h.p.,

$$\begin{aligned}
&\frac{k_{j+1}}{k}\,\mathrm{Pr}_2(S(0:j+1\mid\mathscr{I}_{j+1})\leqslant\eta)\\
&\quad\approx\frac{k_{j+1}}{k}\,\mathrm{Pr}_2(S(0:j\mid\mathscr{I}_{j+1})+S(j+1:j+1\mid\mathscr{I}_{j+1})\leqslant\eta)\\
&\quad\approx(k_{j+1}/k)\,\mathrm{Pr}_3(Z_1+Z_2\leqslant\eta),
\end{aligned}$$

where Z_1 and Z_2 are independent r.v.'s defined over a probability space Ω_3 with probability measure $\mathrm{Pr}_3(\cdot)$, and where $Z_2\sim N(\gamma_{j+1}n_{j+1},n_{j+1})$ and Z_1 has c.d.f.

$$\begin{aligned}
F_{Z_1}(\zeta)&:=\mathrm{Pr}_2(S(0:j\mid\mathscr{I}_{j+1})\leqslant\zeta)\\
&\leqslant\frac{\mathrm{Pr}_2(S(0:j)\leqslant\zeta)}{|\mathscr{I}_{j+1}|/k}\\
&=\frac{k}{k_{j+1}}\,\mathrm{Pr}_2(S(0:j)\leqslant\zeta)\\
&\lesssim\frac{k}{k_{j+1}}\left(\Phi\left(\frac{\zeta-\mu_{j,1}}{\sigma_{j,1}}\right)+(1+\epsilon)\,\Phi\left(\frac{\zeta-\mu_{j,2}}{\sigma_{j,2}}\right)\right)
\end{aligned}$$

by the induction hypothesis. But now we can write $\mathrm{Pr}_3(Z_1+Z_2\leqslant\eta)$ as a convolution integral. To bound this quantity, we introduce yet another probability space, Ω_4, with independent r.v.'s $Z_1'\sim N(\mu_{j,1},\sigma_{j,1}^2)$, $Z_1''\sim N(\mu_{j,2},\sigma_{j,2}^2)$, and $Z_2'\sim N(\gamma_{j+1}n_{j+1},n_{j+1})$. Since the convolution is linear in the c.d.f. $F_{Z_1}(\cdot)$, we see that (w.h.p.)

$$\begin{aligned}
&\mathrm{Pr}_3(Z_1+Z_2\leqslant\eta)\\
&\quad\lesssim\frac{k}{k_{j+1}}\,(\mathrm{Pr}_4(Z_1'+Z_2'\leqslant\eta)+(1+\epsilon)\,\mathrm{Pr}_4(Z_1''+Z_2'\leqslant\eta)).
\end{aligned}$$

Since the sum of independent normal r.v.'s is also normal, we obtain

$$\begin{aligned}
\mathrm{Pr}_3(Z_1+Z_2\leqslant\eta)\lesssim\frac{k}{k_{j+1}}\Bigg(&\Phi\left(\frac{\eta-(\mu_{j,1}+\gamma_{j+1}n_{j+1})}{\sqrt{\sigma_{j,1}^2+n_{j+1}}}\right)\\
&+(1+\epsilon)\,\Phi\left(\frac{\eta-(\mu_{j,2}+\gamma_{j+1}n_{j+1})}{\sqrt{\sigma_{j,2}^2+n_{j+1}}}\right)\Bigg).
\end{aligned}$$

Combining the upper bounds on the two terms in (5.68), we have (w.h.p.) for all $\eta\leqslant T_{j+1}$

$$\begin{aligned}
&\mathrm{Pr}_2(S(0:j+1)\leqslant\eta)\\
&\quad\lesssim\Phi\left(\frac{\eta-T_j}{\sqrt{n_{j+1}}}\right)+\Phi\left(\frac{\eta-(\mu_{j,1}+\gamma_{j+1}n_{j+1})}{\sqrt{\sigma_{j,1}^2+n_{j+1}}}\right)\\
&\qquad+(1+\epsilon)\,\Phi\left(\frac{\eta-(\mu_{j,2}+\gamma_{j+1}n_{j+1})}{\sqrt{\sigma_{j,2}^2+n_{j+1}}}\right).\qquad(5.69)
\end{aligned}$$

Using the definitions of k_s, n_s, and γ_s, we see that $\gamma_s n_s$ grows very rapidly with s, and hence the second and third terms above on the right side of the inequality each correspond to normal distributions with mean approximately $\gamma_{j+1}n_{j+1}\approx\sqrt{2/\pi}\,n_{j+1}/\sqrt{k_{j+1}}$. However, the middle distribution has variance approximately n_j, while the last distribution has variance approximately n ($\gg n_j$ for $j\geqslant 1$). Thus for $j\geqslant 1$ (i.e., for $j+1\geqslant 2$), since η is the left tail of both distribution ($\eta\leqslant T_{j+1}\ll\sqrt{2/\pi}\,n_{j+1}/\sqrt{k_{j+1}}$) and $n_j\ll n$, it follows that the middle term in (5.69) is negligible in comparison with the last term. (We can absorb the middle term into the final term by shifting the mean of the final distribution very slightly.)

When $j=0$, the distributions corresponding to the second and third terms on the right in (5.69) both have variance approximately $n_0+n_1\approx n$, but the middle distribution has mean approximately $(2+40)\sqrt{n}$, while the final distribution has mean approximately $(11+40)\sqrt{n}$. Since η is in the left tail of both distributions, it follows that the last term is a small fraction of the middle term. Multiplying the middle term by $1+\epsilon$ for a small (but not asymptotically vanishing) constant ϵ allows us to discard the last term, and thus we once again obtain (w.h.p.) an upper bound

$$\begin{aligned}
&\mathrm{Pr}_2(S(0:j+1)\leqslant\eta)\\
&\quad\lesssim\Phi\left(\frac{\eta-T_j}{\sqrt{n_{j+1}}}\right)+(1+\epsilon)\,\Phi\left(\frac{n-\gamma_{j+1}n_{j+1}}{\sqrt{\sum_{i=0}^{j+1}n_i}}\right),
\end{aligned}$$

which agrees with (5.53) for $s=j+1$. To complete the induction step, we set $\eta=T_{j+1}$ and substitute for the different variables. We find that, w.h.p.,

$$\mathrm{Pr}_2(S(0:j+1)\leqslant T_{j+1})\leqslant\frac{k_{j+2}}{k}$$

$$\text{for all}\quad j\in\{0,1,\ldots,N-2\},$$

as required.

After the final step (Step N), we use a somewhat different technique (described after Lemma 5.17) to show that w.h.p. $\mathrm{Pr}_2(S(0:N)\leqslant 0)=0$, as required. For the CLT (actually just a souped-up DeMoivre–Laplace theorem) to hold at each step, it suffices to have (for $1\leqslant j\leqslant N$)

$$T_j\ll(n_j)^{2/3}\qquad\text{and}\qquad\frac{n_j}{\sqrt{k_j}}\ll n^{2/3}.$$

Now $T_j=\sqrt{n}/2^{j-1}$, $n_j\approx n/10^{2^j}$, and $k_j\approx n/+1000^{2^j}$, where $10^{2^j}\leqslant 10^{2^N}\leqslant n^{0.01}$. It follows readily that the desired inequalities hold. This completes the extended plausibility argument for the truth of Theorem 3.2.

5.3. Lemmas for Theorem 3.2

In this subsection we introduce a number of lemmas from which Theorem 1.4 will follow. The next subsections are for the proofs of (some of) the lemmas.

We begin by defining the phrase "with high probability."

DEFINITION. As usual, let n be the total number of columns in the matrix X. We say that an event $\mathscr{E}$ (more precisely, a family of events $\mathscr{E}_n$ indexed by n) occurs *with high probability* if, for every $\alpha > 0$, there exists $n_0 = n_0(\alpha)$ such that

$$\Pr(\mathscr{E}_n) \geqslant 1 - n^{-\alpha} \qquad \text{for all} \quad n \geqslant n_0(\alpha).$$

Remark. Let $t := t(n)$ be any integer-valued function of n such that $n^{1/2} \leqslant t \leqslant n$ for all n. Then $\mathscr{E}$ occurs w.h.p. if $\forall \alpha > 0$, $\exists t_0(\alpha)$ such that $\forall t \geqslant t_0(\alpha)$,

$$\Pr(\mathscr{E}_n) \geqslant 1 - t^{-\alpha}.$$

In practice, we shall have t an element of $\{k_s, k - k_s, n_s\}_{1 \leqslant s \leqslant N}$ or a sum of elements from this set, and t will indeed satisfy $n^{1/2} \leqslant t \leqslant n$.

There will be only finitely many lemmas, each possibly with its own function $t_0(\alpha)$, so we can get a uniform definition of "w.h.p." by taking the function $n_0(\alpha)$ to be the maximum of the lemma-specific functions $n_0(\alpha)$. Since each lemma will be applied only polynomially many times (in n), it follows from the union bound on probabilities that our conclusions also hold w.h.p.

DEFINITION. As usual, given any r.v. ξ, its c.d.f. is the right-continuous function F_ξ such that

$$F_\xi(u) = \Pr(\xi \leqslant u) \qquad \text{for all} \quad u \in \mathbb{R}.$$

DEFINITION. Given any collection of r.v.'s $\xi_1, ..., \xi_t$ (not necessarily independent or identically distributed), given a *realization* $\hat{\xi}_1, ..., \hat{\xi}_t$ of the r.v.'s, the collection's *sample c.d.f.*, or *empirical c.d.f.*, is the right-continuous function $\hat{F}_\xi$ such that

$$\hat{F}_\xi(u) = t^{-1} \, |\{i : \hat{\xi} \leqslant u\}| \quad \text{for all } u \in \mathbb{R}.$$

The first lemma below will imply that on any given Step s ($0 \leqslant s \leqslant N$), the r.v.'s $w_{ij} := X_{ij} z_j$ may be approximated by i.i.d. r.v.'s $y_{ij} \sim \text{Bern}((1 + \gamma_s)/2)$, provided that we restrict our attention to a sufficiently small number of rows (i.e., values of i).

LEMMA 5.1. *Suppose that t is odd and b is even, with $b \leqslant t^{1/10}$. Let $\xi_1, ..., \xi_t$ be i.i.d. r.v.'s $\sim$ Bern(1/2). (That is, $\Pr(\xi_i = 1) = \Pr(\xi_i = -1) = 1/2$ for each i.) Let $\mathscr{B} = \{m_1, ..., m_b\}$ be any b-element subset of $\{1, ..., t\}$. Then there exists an absolute constant t_0 (independent of b, in particular) such that for all $t \geqslant t_0$, we can define mutually i.i.d. r.v.'s $\psi_1, ..., \psi_b \sim$ Bern$(\frac{1}{2}(1 + \sqrt{2/(\pi t)}(1 - t^{-1/8})))$ on the same space as the ξ's such that*

$$\psi_j \leqslant \xi_{m_j} \operatorname{sgn}\left(\sum_{i=1}^{t} \xi_i\right) \qquad \text{for all} \quad j \in [b].$$

Proof. Given in Subsection 5.4.

Using Lemma 5.1 on Step s and letting $t = k_s$, we can approximate the joint distribution of the partial inner products $\{S^{(r)}(s : s)\}_{r \in \mathscr{B}}$ for any $\mathscr{B} \subseteq \mathscr{I}_s$ satisfying $b := |\mathscr{B}| \leqslant t^{1/10}$. We obtain b i.i.d. r.v.'s, each the sum of n_s i.i.d. Bernoulli r.v.'s with parameter $(1 + \gamma_s)/2$, where

$$\gamma_s := \sqrt{2/(\pi t)}\,(1 - t^{-1/8}).$$

Thus, approximately, each r.v. $S^{(r)}(s : s)$ has binomial distribution. (More precisely, each $S^{(r)}(s : s)$ is at least as great as a r.v. having the appropriate binomial distribution.) For any subset of row indices $\mathscr{A} \subseteq \bar{\mathscr{I}}_s$, the situation is similar but even simpler. In this case, the partial inner products $\{S^{(r)}(s : s)\}_{r \in \mathscr{A}}$ are exactly i.i.d. binomial r.v.'s, each the sum of n_s i.i.d. Bern (1/2) r.v.'s.

Whether the rows have indices in $\mathscr{I}_s$ or in $\bar{\mathscr{I}}_s$, it will be useful to bound the appropriate c.d.f. by a normal c.d.f. This we do in the following lemmas.

DEFINITION (See [32, pp. 73–75]). For j, m integers, $p \in [0, 1]$, with $0 \leqslant j \leqslant m$, let

$$b(m, p, j) := \binom{m}{j} p^j (1 - p)^{m - j} \tag{5.70}$$

and

$$B(m, p, j) := \sum_{i=0}^{j} b(m, p, i). \tag{5.71}$$

Thus $b(\cdot, \cdot, \cdot)$ is the probability mass function (p.m.f.) of a binomial r.v., and $B(\cdot, \cdot, \cdot)$ is its (discrete) c.d.f.

Remark. A r.v. ξ with distribution as above is the sum of m independent Bernoulli r.v.'s each taking a value in $\{0, 1\}$. Sometimes we also refer to the r.v. $2\xi - m$, which is the sum of independent Bernoulli r.v.'s in $\{-1, 1\}$, as a binomially distributed r.v. The exact definition should be clear from the context.

LEMMA 5.2 (Local Limit Lemma; based on [32, pp. 73–75]). *Let j, m be integers, $p \in (0, 1)$, and fix any ϵ_1 with $0 < \epsilon_1 < 1/6$. Suppose that for each sufficiently large m we have*

$$m^{-1/3 + \epsilon_1} < p < 1 - m^{-1/3 + \epsilon_1}$$

and

$$|j-mp| \leqslant (mp(1-p))^{2/3-\epsilon_1}.$$

Then as $m \to \infty$,

$$b(m, p, j) \sim (2\pi mp(1-p))^{-1/2} \times \exp(-(j-mp)^2/2mp(1-p)), \qquad (5.72)$$

and the convergence of the ratio of the two sides to 1 *is* uniform *for* p *and* j *in the given ranges. (The speed of convergence depends only on* ϵ_1.)

Proof. The lemma follows essentially as in [32, pp. 73–75], as an application of Stirling's approximation. The bounds on p and $|j-mp|$ ensure that $|j-mp|/mp$, $|j-mp|/m(1-p)$, and $|j-mp|^3 (p^{-2}+(1-p)^{-2})/m^2$ all converge uniformly to 0 as $m \to \infty$, as required in the Taylor series approximation of the quantity

$$-j \log(1+(j-mp)/mp) - (m-j)\log(1-(j-mp)/m(1-p)),$$

which arises from Stirling's approximation for $\log(b(m, p, j))$. ∎

LEMMA 5.3. *Fix any* ϵ_1 *with* $0 < \epsilon_1 < 1/6$. *Suppose, as in the previous lemma, that*

$$m^{-1/3+\epsilon_1} < p < 1-m^{-1/3+\epsilon_1},$$

but now suppose that

$$x := (mp(1-p))^{2/3-\epsilon_1}$$

is a lower *bound on* $|j-mp|$. *Then there exists* $M = M(\epsilon_1)$ *such that for all* $m > M(\epsilon_1)$,

$$\sum_{j \geqslant mp+x} b(m, p, j) \leqslant \exp(-m^{(4/3)(1/6-\epsilon_1)}) \qquad (5.73)$$

and

$$\sum_{j \leqslant mp-x} b(m, p, j) \leqslant \exp(-m^{(4/3)(1/6-\epsilon_1)}). \qquad (5.74)$$

(*Note that* $M(\epsilon_1)$ *does not depend on* p *or* x.)

Proof. Given in Subsection 5.5.

LEMMA 5.4. *Fix any* $\epsilon > 0$. *Then there exists* $m_0 = m_0(\epsilon)$ (*not depending on* p *or* j) *such that for all* $m \geqslant m_0$, *for all* $j \in \{0, 1, ..., m\}$, *for all* $p \in [1/10, 9/10]$,

$$B(m, p, j) \leqslant \exp(-m^{1/6}) + (1+\epsilon)\,\Phi((j-mp)/(mp(1-p))^{1/2}). \qquad (5.75)$$

Proof. Given in Subsection 5.5.

In the next lemma, the r.v.'s $\xi_1, ..., \xi_t$ may be thought of as indicator r.v.'s. By letting $\xi_i = I\{S^{(r_i)}(s:s) \leqslant m\}$ for arbitrary but fixed $m \in \{-n_s, -n_s+2, ..., n_s-2, n_s\}$ with $t = k_s$, we shall be able to conclude that w.h.p. the sample c.d.f. of $\{S^{(r)}(s:s)\}_{r \in \mathscr{F}_s}$ is bounded above by a slightly modified normal c.d.f. The lemma will also imply a similar result about the sample c.d.f. of $\{S^{(r)}(s:s)\}_{r \in \bar{\mathscr{F}}_s}$, but we shall not need this second result.

LEMMA 5.5. *Suppose that* $\xi_1, ..., \xi_t$ *are exchangeable* $0-1$ *r.v.'s. (That is, the* ξ_i*'s are identically distributed and take values in* $\{0, 1\}$, *and the joint distribution of* $\xi_1, ..., \xi_t$ *is the same as that of* $\xi_{\pi(1)}, ..., \xi_{\pi(t)}$ *for every permutation* π *on* $\{1, ..., t\}$.)

Let

$$b := 2\lfloor (1/2)\, t^{1/10} \rfloor, \qquad (5.76)$$

and suppose that

$$\Pr(\xi_1 = \xi_2 = \cdots = \xi_b = 1) \leqslant q^b, \qquad \text{where} \quad q \in (0, 1).$$

Then there exists an absolute constant t_0 *such that for all* $t \geqslant t_0$,

$$\Pr\left(\sum_{i=1}^{t} \xi_i \geqslant t^{3/5} + (1+t^{-1/12})\, qt\right) \leqslant \exp(-t^{1/70}). \qquad (5.77)$$

(*By an absolute constant, we mean that* t_0 *does not depend on the value of* q *or on the distribution of the* ξ*'s.*)

Proof. Given in Subsection 5.6.

The next two lemmas and the intervening corollary will allow us to move back and forth between bounds on statistical c.d.f.'s and bounds on sample c.d.f.'s.

LEMMA 5.6. *Suppose that* $\psi_1, ..., \psi_t$ *are real-valued exchangeable r.v.'s each drawn from the same set of size (at most)* M. (M *can be a constant or can depend arbitrarily on* t.) *Also suppose*

$$\Pr(\psi_1 \leqslant u, ..., \psi_b \leqslant u) \leqslant [G(u)]^b \qquad \text{for all} \quad u \in \mathbb{R},$$

where $b = 2\lfloor (1/2)\, t^{1/10} \rfloor$ *and* $G(\cdot)$ *is a nonnegative and nondecreasing real-valued function.*

Let $\hat{F}(\cdot)$ *be the sample c.d.f. of the* ψ*'s. Then there exists an absolute constant* t_0 *such that for all* $t \geqslant t_0$,

$$\Pr(\hat{F}(u) \leqslant t^{-2/5} + (1+t^{-1/12})\, G(u)\ \forall u \in \mathbb{R}) \geqslant 1 - M \cdot \exp(-t^{1/70}).$$

Proof. The lemma follows readily from Lemma 5.5 and the union bound upon letting $\xi_i := I\{\psi_i \leqslant u\}$, where u takes on each of the M possible values of the ψ's in turn. ▮

COROLLARY 5.1. *Suppose that $\psi_1, ..., \psi_t$ are real-valued i.i.d. r.v.'s with common c.d.f. F, with all ψ_i's drawn from a set of size (at most) M, and suppose that $F(u) \leqslant G(u)$ for all $u \in \mathbb{R}$. Let $\hat{F}(\cdot)$ be the sample c.d.f. of the ψ's. Then there is an absolute constant t_0 such that for all $t \geqslant t_0$,*

$$\Pr(\hat{F}(u) \leqslant t^{-2/5} + (1 + t^{-1/12})\, G(u)\ \forall u \in \mathbb{R}) \geqslant 1 - M \cdot \exp(-t^{1/70}).$$

LEMMA 5.7. *Given a collection of t real numbers $c_1 \leqslant \cdots \leqslant c_t$ (not necessarily distinct), define the corresponding empirical c.d.f. $\hat{F}$ in the usual way. Define mutually i.i.d. r.v.'s $\xi_1, ..., \xi_t$, each with statistical c.d.f.*

$$G(u) = \begin{cases} \min\{t^{-1/6} + \hat{F}(u), 1\} & \text{if} \quad u \geqslant c_1 \\ 0 & \text{if} \quad u < c_1, \end{cases} \tag{5.78}$$

and let $\hat{G}(\cdot)$ be the sample *c.d.f. of the ξ's. Then there is an absolute constant t_0 such that for all $t \geqslant t_0$,*

$$\Pr(\hat{G}(u) \geqslant \hat{F}(u)\ \forall u \in \mathbb{R}) \geqslant 1 - \exp(-t^{1/9}). \tag{5.79}$$

Proof. Given in Subsection 5.7.

LEMMA 5.8. *Let $U_1, ..., U_t$ be exchangeable real-valued r.v.'s taking values on a finite set (whose size can grow with t), and suppose that, w.h.p., their sample c.d.f. is dominated by $\hat{F}(\cdot)$. Then we can define i.i.d. r.v.'s $W_1, ..., W_t$ (defined on the same space as the U_i's) with c.d.f. $G(\cdot)$ as in Lemma 5.7, such that, w.h.p.,*

$$U_i \geqslant W_i \qquad \text{for all} \quad i \in \{1, ..., t\}.$$

Proof. Follows from Lemma 5.7, using a construction similar to that of Lemma 5.22. ▮

The next two lemmas are useful when manipulating normal distribution functions.

LEMMA 5.9 (See, e.g., [38, pp. 82–83]). *The normal distribution function*

$$\Phi(-\alpha) := (2\pi)^{-1/2} \int_{-\infty}^{-\alpha} \exp\{-\beta^2/2\}\, d\beta = (2\pi)^{-1/2} \int_{\alpha}^{\infty} \exp(-\beta^2/2)\, d\beta$$

satisfies the bounds

$$(2\pi)^{-1/2}\, \alpha^{-1} \exp(-\alpha^2/2)(1 - \alpha^{-2}) < \Phi(-\alpha) < (2\pi)^{-1/2}\, \alpha^{-1} \exp(-\alpha^2/2) \quad \text{for all} \quad \alpha > 0. \tag{5.80}$$

Proof. The lemma follows readily upon integrating $\beta^{-1}(\beta \exp(-\beta^2/2))$ by parts. ▮

LEMMA 5.10. *If $u \geqslant A \geqslant 4$, then*

$$\Phi(-u) \leqslant e^{1-A} \Phi(-(u-1)). \tag{5.81}$$

Proof. By Lemma 5.9,

$$\begin{aligned} \Phi(-(u-1)) &\geqslant (2\pi)^{-1/2}\, (u-1)^{-1} \\ &\quad \times \exp(-(u^2 - 2u + 1)/2)(1 - (u-1)^{-2}) \\ &\geqslant (2\pi)^{-1/2}\, u^{-1} \exp(-u^2/2) \exp(u - 1/2)(8/9) \\ &\geqslant \Phi(-u) \cdot \exp(A - 1/2 + \log(8/9)) \\ &\geqslant \exp(A-1)\, \Phi(-u). \end{aligned}$$ ▮

Now we combine the preceding lemmas to obtain bounds on the distributions of $S(s : s \mid \mathscr{I}_s)$ and $S(s : s \mid \bar{\mathscr{I}}_s)$, the partial inner products from Step s contributed by the "voting" rows and the "nonvoting" rows, respectively.

LEMMA 5.11. *On Step s of the procedure for constructing the vector z (for $0 \leqslant s \leqslant N$), w.h.p. the sample c.d.f. of $\{S^{(r)}(s : s)\}_{r \in \mathscr{I}_s}$ satisfies*

$$\Pr_2(S(s : s \mid \mathscr{I}_s) \leqslant \eta) \leqslant (9/8)\, k_s^{-2/5} + (5/4)\, \Phi((\eta - \mu(s : s))/\sigma(s : s)),$$

where $t = k_s$,

$$\mu(s : s) = \gamma_s n_s = (\sqrt{2/(\pi t)})(1 - t^{-1/8})\, n_s,$$

and

$$\sigma^2(s : s) = n_s(1 - \gamma_s^2) = n_s[1 - (2/(\pi t))(1 - t^{-1/8})^2].$$

Proof. Let $t = k_s$, and let $r_1, ..., r_t$ be the t elements of $\mathscr{I}_s$. For $1 \leqslant i \leqslant t$, $j \in \mathscr{I}_s$, let

$$v_{ij} := X_j^{(r_i)} \cdot \operatorname{sgn}\left(\sum_{m=1}^{t} X_j^{(r_m)} \right).$$

By Lemma 5.1, for $\mathscr{B} = \{m_1, ..., m_b\}$ any b-element subset of $\{1, ..., t\}$, where $b = 2 \lfloor (1/2)\, t^{1/10} \rfloor \approx t^{1/10}$, there exists an absolute constant t_0 such that for all $t \geqslant t_0$, we can define

mutually i.i.d. r.v.'s $\{y_{ij}\}_{i\in\mathscr{B},\, j\in\mathscr{J}_s}$ on the same space as the v_{ij}'s, such that

$$y_{ij} \sim \text{Bern}((1+\gamma_s)/2)$$

and

$$y_{ij} \leqslant v_{ij} \quad \text{for all } i\in\mathscr{B},\ j\in\mathscr{J}_s.$$

Now, by Lemma 5.4, with $m=n_s$, $j=(\eta+n_s)/2$, $\epsilon=1/8$, and $p=(1+\gamma_s)/2$, we see that for all sufficiently large n, for $\mathscr{B}$ any b-element subset of $\{1, ..., t\}$, for all $\eta\in\mathbb{R}$,

$$\begin{aligned}
&(\Pr(S^{(r_i)}(s:s)\leqslant\eta\ \forall i\in\mathscr{B}))^{1/b}\\
&\quad=\left(\Pr\left(\sum_{j\in\mathscr{J}_s} v_{ij}\leqslant\eta\ \forall i\in\mathscr{B}\right)\right)^{1/b}\\
&\quad\leqslant\Pr\left(\sum_{j\in\mathscr{J}_s} y_{ij}\leqslant\eta\right)\\
&\quad\leqslant\exp(-n_s^{1/6})+(9/8)\,\Phi\left(\frac{(\eta-\gamma_s n_s)/2}{((n_s/4)(1-\gamma_s^2))^{1/2}}\right)\\
&\quad=\exp(-n_s^{1/6})+(9/8)\,\Phi\left(\frac{\eta-\gamma_s n_s}{(n_s(1-\gamma_s^2))^{1/2}}\right).
\end{aligned}$$

Now by Lemma 5.6, for all sufficiently large n, w.h.p. the *sample* distribution of $\{S^{(r_i)}(s:s)\}_{r_i\in\mathscr{I}_s}$ satisfies

$$\begin{aligned}
&\Pr_2(S(s:s\mid\mathscr{I}_s)\leqslant\eta)\\
&\quad\leqslant t^{-2/5}+(1+t^{-1/12})\\
&\qquad\times\left(\exp(-n_s^{1/6})+(9/8)\,\Phi\left(\frac{\eta-\gamma_s n_s}{\sqrt{n_s(1-\gamma_s^2)}}\right)\right)\\
&\quad\leqslant(9/8)\,k_s^{-2/5}+(5/4)\,\Phi\left(\frac{\eta-\gamma_s n_s}{\sqrt{n_s(1-\gamma_s^2)}}\right)
\end{aligned}$$

for all $\eta\in\mathbb{R}$. ∎

LEMMA 5.12. *Letting t, $\mu(s:s)$, and $\sigma(s:s)$ be as in the previous lemma, we define mutually i.i.d. r.v.'s (actually* extended *r.v.'s, since they can take on the value $-\infty$) $\xi_1, ..., \xi_t$, each with statistical c.d.f.*

$$G(\eta)=\min\left\{1,\,(5/4)\,t^{-1/6}+(5/4)\,\Phi\left(\frac{\eta-\mu(s:s)}{\sigma(s:s)}\right)\right\}.$$

Then w.h.p., the sample *c.d.f. $\hat{G}(\cdot)$ of the ξ_i's satisfies*

$$\hat{G}(\eta)\geqslant\Pr_2(S(s:s)\mid\mathscr{I}_s)\leqslant\eta \quad \text{for all } \eta\in\mathbb{R}.$$

Furthermore, since the $\{S^{(r)}(s:s)\}$ are exchangeable, we can define i.i.d. r.v.'s $\xi_1, ..., \xi_t$ as above on the same space as the $\{S^{(r)}(s:s)\}$, such that w.h.p.

$$\xi_i\leqslant S^{(r_i)}(s:s) \quad \text{for all } i\in\{1, ..., t\}.$$

Proof. The proof follows readily from Lemmas 5.11, 5.7, and 5.8. ∎

Now we give a similar (but slightly simpler) lemma dealing with the c.d.f. of $S(s:s\mid\bar{\mathscr{I}}_s)$.

LEMMA 5.13. *On Step s of the procedure for constructing the vector z (where $0\leqslant s\leqslant N$), let $t=k-k_s$ and $\{r_1, ..., r_t\}=\bar{\mathscr{I}}_s$. Then we can define mutually i.i.d. extended r.v.'s $\xi_1, ..., \xi_t$, each with c.d.f.*

$$G(\eta)=\min\{1,\,(5/4)\,t^{-1/6}+(5/4)\,\Phi(\eta/n_s^{1/2})\},$$

on the same space as $\{S^{(r)}(s:s)\}_{r\in\bar{\mathscr{I}}_s}$, such that

$$S^{(r_i)}(s:s)\geqslant\xi_i \qquad \text{for all} \quad i\in[t].$$

Proof. For $i\in[t]$, $j\in\mathscr{J}_s$, let

$$\bar{v}_{ij}:=X_j^{(r_i)}\cdot\operatorname{sgn}\left(\sum_{r\in\mathscr{I}_s} X_j^{(r)}\right).$$

Then the r.v.'s $\{\bar{v}_{ij}\}_{i\in[t],\, j\in\mathscr{J}_s}$ are mutually i.i.d. $\sim\text{Bern}(1/2)$. Thus the values $S^{(r_i)}(s:s)$ are i.i.d. symmetric binomial r.v.'s. The lemma follows readily from Lemma 5.4 with $m=n_s$, $j=(\eta+n_s)/2$, $\epsilon=1/8$, and $p=1/2$, together with the fact that

$$\exp(-n_s^{1/6})\ll(k-k_s)^{-1/6}.$$ ∎

Assuming the truth of the lemmas whose proofs have been deferred, we can now state and prove a precise version of the bound on the c.d.f. of $S(0:s)$ given in (5.53) in Subsection 5.2, at least for the case $s=0$.

LEMMA 5.14. *After Step 0 of the construction of the vector z (described in Subsection* 5.1*), we have w.h.p.,*

$$\begin{aligned}
&\Pr_2(S(0:0)\leqslant\eta)\\
&\quad\leqslant(5/4)\,k^{-2/5}+(5/4)\,\Phi\left(\frac{\eta-\mu_{0,2}}{\sigma_{0,2}}\right) \qquad \text{for all} \quad \eta\leqslant T_0,
\end{aligned}\tag{5.82}$$

where

$$\mu_{0,2}=\gamma_0 n_0(\approx\sqrt{2/\pi k}\,n_0\approx 11.2\sqrt{n})$$

and

$$\sigma_{0,2}^2=n_0(\approx n).$$

Furthermore, w.h.p.,

$$\Pr_2(S(0:0) \leqslant T_0) := \Pr_2(S(0:0) \leqslant 2n^{1/2}) \leqslant k_1/k (\approx 2 \times 10^{-6}). \tag{5.83}$$

Proof. In Step 0, $k_0 = k$; i.e., all rows of X are allowed to vote on the first n_0 components of z. The main part of the lemma then follows readily from Lemma 5.11. Since $\eta \leqslant T_0 < \mu_{0,2}$, we may take $\sigma_{0,2}^2 = n_0 (> n_0(1-\gamma_0^2))$. The bound on $\Pr(S(0:0) \leqslant T_0)$ follows readily from Lemma 5.9. ∎

In using induction to establish the probabilistic upper bound on the c.d.f. of $S(0:s)$ given in (5.53), we shall need the following simple lemma.

LEMMA 5.15. *Let m and t be integers with $1 \leqslant t \leqslant m$, let T be some threshold, and let $\bar{F}(\cdot)$ be a nonnegative nondecreasing function from $\mathbb{R}$ to $\mathbb{R}$. Suppose that the collection $\{\hat{\xi}_1, ..., \hat{\xi}_m\}$ has empirical c.d.f. $\hat{F}(\cdot)$, where $\hat{F}(u) \leqslant \bar{F}(u)$ for all $u \in \mathbb{R}$ and $\bar{F}(T) \leqslant t/m$.*

Now permute the $\hat{\xi}_i$'s to put them in increasing order: $\hat{\xi}^{(1)} \leqslant \cdots \leqslant \hat{\xi}^{(m)}$. Then for all $t \in [m]$ such that $\bar{F}(T) \leqslant t/m$, the sample c.d.f. $\hat{G}_t(\cdot)$ of $\{\hat{\xi}^{(1)}, ..., \hat{\xi}^{(t)}\}$ satisfies

$$\hat{G}_t(u) \leqslant \frac{m}{t} \bar{F}(u) \qquad \text{for all} \quad u \in \mathbb{R}, \tag{5.84}$$

and the upper bound is itself bounded above by 1 for all $u \leqslant T$.

Proof. Straightforward. ∎

LEMMA 5.16. *Suppose that ξ and ξ' are r.v.'s with respective c.d.f.'s F_ξ and $F_{\xi'}$ such that*

$$F_\xi \leqslant a + bF_{\xi'}, \tag{5.85}$$

where a and b are nonnegative constants. Similarly, suppose that ψ and ψ' are r.v.'s satisfying

$$F_\psi \leqslant c + dF_{\psi'}, \tag{5.86}$$

where c and d are nonnegative constants. If ξ and ψ are independent, as are ξ' and ψ', then

$$F_{\xi+\psi} \leqslant (c + ad) + (bd)\, F_{\xi'+\psi'} \tag{5.87}$$

and

$$F_{\xi+\psi} \leqslant (a + bc) + (bd)\, F_{\xi'+\psi'}. \tag{5.88}$$

Proof. We might as well assume that ξ, ξ', ψ, and ψ' are all mutually independent. Then, by the nonnegativity and linearity of the convolution integral defining the c.d.f. of a sum of independent r.v.'s,

$$F_{\xi+\psi'} \leqslant a + bF_{\xi'+\psi'}$$

and

$$F_{\xi+\psi} \leqslant c + dF_{\xi+\psi'}.$$

Thus

$$F_{\xi+\psi} \leqslant c + d(a + bF_{\xi'+\psi'}) = (ad + c) + (bd)\, F_{\xi'+\psi'}.$$

The other bound on $F_{\xi+\psi}$ follows in the same way. ∎

Now we state the main lemma from which Theorem 1.4 will follow almost immediately.

LEMMA 5.17. *W.h.p.,*

$$\Pr_2(S(0:s) \leqslant \eta) \leqslant 3^s n^{-1/10} + \Phi\left(\frac{\eta - \mu_{s,1}}{\sigma_{s,1}}\right) + (3/2)(1 + n^{-1/10})^s\, \Phi\left(\frac{\eta - \mu_{s,2}}{\sigma_{s,2}}\right) \tag{5.89}$$

for all $\eta \leqslant T_s$, for all $s \in \{0, 1, ..., N\}$, where

$$\mu_{s,1} = T_{s-1}, \quad \sigma_{s,1}^2 = n_s, \quad \mu_{s,2} = n_s \gamma_s \approx \sqrt{2/\pi}\, n_s/\sqrt{k_s}, \quad \text{and}$$

$$\sigma_{s,2}^2 = \sum_{j=0}^{s} n_j \approx n.$$

Furthermore, w.h.p.,

$$\Pr_2(S(0:s) \leqslant T_s) \leqslant k_{s+1}/k, \qquad \forall s \in \{0, 1, ..., N-1\}.$$

Remark. The last inequality of the lemma is pivotal in establishing the lemma by induction.

Proof of Lemma 5.17. The proof proceeds by induction essentially as in the proof sketch given in Section 5.2. The only real differences are that we keep track more carefully of the constants on the right-hand side of the main inequality and hat we justify statements about quasi-independence and quasi-normality using the lemmas of this subsection.

In order to establish the induction, we follow Steps 0 through 6 below.

Step 0: Let $t = k_{s+1}$ and write $\mathscr{I}_{s+1} = \{r_1, ..., r_t\}$.

Step 1: By the induction hypothesis and Lemma 5.15 (with $m = k$, $t = k_{s+1}$, and $T = T_s$), w.h.p. the sample c.d.f. of $\{S^{(r)}(0:s)\}_{r \in \mathscr{I}_{s+1}}$ is dominated by the scaled left tail of a normal c.d.f. Then by Lemmas 5.7 and 5.8, we can define

i.i.d. r.v.'s $W_1, ..., W_t$ on the same probability space as $\{S^{(r)}(0:s)\}_{r\in\mathscr{I}_{s+1}}$ such that the common distribution of the $\{W_i\}$ is essentially the scaled normal distribution already derived, and w.h.p.,

$$W_i \leqslant S^{(r_i)}(0:s) \qquad \text{for all} \quad i\in\{1, ..., t\}.$$

Step 2: Conditioned on knowledge of the row indices r_i comprising the set $\mathscr{I}_{s+1}$, the r.v.'s $\{S^{(r)}(s+1:s+1)\}_{r\in\mathscr{I}_{s+1}}$ are independent of the r.v.'s $\{S^{(r)}(0:s)\}_{r\in\mathscr{I}_{s+1}}$. Furthermore, by Lemma 5.12, we can define i.i.d. r.v.'s $\xi_1, ..., \xi_t$ (also independent of $\{S^{(r)}(0:s)\}$) having approximately normal distribution on the same space as $\{S^{(r)}(s+1:s+1)\}_{r\in\mathscr{I}_{s+1}}$ such that, w.h.p.,

$$\xi_i \leqslant S^{(r_i)}(s+1:s+1) \qquad \text{for all} \quad i\in\{1, ..., t\}.$$

Step 3: Combining Steps 1 and 2, we can form the t i.i.d. r.v.'s $\{W_i+\xi_i\}_{i\in[t]}$ to get (w.h.p.) lower bounds on the r.v.'s $\{S^{(r)}(0:s+1)\}_{r\in\mathscr{I}_{s+1}}$. Using Lemma 5.16, we are able to handle the scale factors that appear in front of the different normal c.d.f.'s.

Step 4: By Corollary 5.1, w.h.p. we can convert the bound on the statistical c.d.f. of each r.v. $W_i+\xi_i$ into an upper bound on the *empirical* c.d.f. of $\{S^{(r)}(0:s+1)\}_{r\in\mathscr{I}_{s+1}}$.

Step 5: For the rows not already considered, namely, for $r\in\bar{\mathscr{I}}_{s+1}$, the induction hypothesis asserts that w.h.p.

$$S^{(r)}(0:s) \geqslant T_s \quad \text{for all} \quad r\in\bar{\mathscr{I}}_{s+1}.$$

Now conditional upon knowledge of the indices $r\in\bar{\mathscr{I}}_{s+1}$, the r.v.'s $S^{(r)}(s+1:s+1)$ are all independent of each other and of $\{S^{(r)}(0:s)\}$; by DeMoivre–Laplace, the former r.v.'s have approximately normal distribution with zero mean. Thus we can (w.h.p.) lower-bound the r.v.'s $\{S^{(r)}(0:s+1)\}_{r\in\bar{\mathscr{I}}_{s+1}}$ by a set of i.i.d., approximately normal r.v.'s with mean T_s and variance n_{s+1}. Now by Corollary 5.1, w.h.p. we can turn the bound on statistical c.d.f.'s into an upper bound on the *empirical* c.d.f. of $\{S^{(r)}(0:s+1)\}_{r\in\bar{\mathscr{I}}_{s+1}}$.

Step 6: Taking a convex combination of the bounds from Steps 4 and 5, we obtain (w.h.p.) an upper bound on the empirical c.d.f. of all k values $\{S^{(r)}(0:s+1)\}_{r\in[k]}$, completing the induction.

Lemmas 5.9 and 5.10 are used throughout the entire procedure to simplify expressions involving linear combinations of normal c.d.f.'s. The main simplification, which is applied repeatedly, is that of bounding a linear combination of two normal c.d.f.'s (for all arguments in a given semi-infinite interval) by a single scaled normal c.d.f.

For $s=N$, (5.89) goes through as for $s\leqslant N-1$. However, in order to show that $\Pr_2(S(0:N)\leqslant 0)=0$ w.h.p., which is the essence of Theorem 1.4, we need a slightly different approach on Step N. We show in the next lemma and corollary that, w.h.p., none of the partial row sums after Step $N-1$ is less than $-n^{1/2+\epsilon}$ (for arbitrarily small but fixed ϵ); then w.h.p., each $S^{(r)}(N:N)$ for $r\in\mathscr{I}_N$ will be large enough to make $S^{(r)}(0:N-1)+S^{(r)}(N:N)>0$. The rows $r\in\bar{\mathscr{I}}_N$ will, w.h.p., have $S^{(r)}(0:N-1)\geqslant T_{n-1}$, and then the usual application of our modified DeMoivre–Laplace limit theorem shows that, w.h.p., all such rows have $S^{(r)}(N:N) > -T_{N-1}$, as required so that $S^{(r)}(0:N)>0$. ∎

LEMMA 5.18. *For each $r\in\{1, ..., k\}$ and every ϵ in $(0, 1/6)$, there exists $n_0(\epsilon)$ such that for all $n\geqslant n_0(\epsilon)$,*

$$\Pr(S^{(r)}(0:N-1)\leqslant -n^{1/2+\epsilon})\leqslant \exp(-n^{\epsilon}).$$

Proof. It is clear from the construction of the vector z that each partial inner product $S^{(r)}(0:N-1)$ is at least as great as a sum of $\sum_{j=0}^{N-1} n_j$ i.i.d. Bern $(1/2)$ r.v.'s. The lemma then follows from Lemma 5.3, together with the fact that $m=\sum_{j=0}^{N-1} n_j \approx n$. ∎

COROLLARY 5.2. *For every ϵ in $(0, 1/6)$, there exists $n_0(\epsilon)$ such that for all $n\geqslant n_0$,*

$$\Pr(S^{(r)}(0:N-1) > -n^{1/2+\epsilon}, \forall r\in[k]) \geqslant 1-k\exp(-n^{\epsilon}) \geqslant 1-n\exp(-n^{\epsilon}).$$

Taking $\epsilon=0.001$, say, in the corollary above, and using the fact that for any individual $r\in\mathscr{I}_N$, we have $S^{(r)}(N{:}\ N)\approx N(\gamma_N n_N, n_N)$, where $\gamma_N n_N=\Omega(n_N/\sqrt{k_N})>n^{0.502}\gg n^{1/2+\epsilon}$ and $\gamma_N n_N/\sqrt{n_N}=\Omega(\sqrt{n_N/k_N})>n^{0.002}\gg n^{\epsilon}$. Then by Lemma 5.4, for each $r\in\mathscr{I}_N$,

$$\Pr(S^{(r)}(N:N) > -S^{(r)}(0:N-1)) = 1-O(n^{-\alpha})$$

for every $\alpha>0$; by the union bound, the result holds w.h.p. for all k rows simultaneously, and Theorem 1.4 is proved (modulo the lemmas whose proofs were deferred).

5.4. *Proof of Lemma* 5.1

In proving Lemma 5.1, we can assume without loss of generality that $\mathscr{B}=\{1, ..., b\}$. We begin by proving the following lemma.

LEMMA 5.19. *For odd t, suppose that $\xi_1, ..., \xi_t$ are i.i.d. r.v.'s with Bern $(1/2)$ distribution, i.e.,*

$$\Pr(\xi_i=1)=\Pr(\xi_i=-1)=1/2.$$

Let $\xi_{maj}:=\operatorname{sgn}(\sum_{i=1}^{t}\xi_i)$, and for $1\leqslant i\leqslant t$, let $\theta_i:=\xi_i\xi_{maj}$. Let b be any even positive integer with $b\leqslant t^{1/10}$, and let

$$N_\theta(b) := |\{i\in\{1, ..., b\}: \theta_i=1\}|.$$

(Note that $N_\theta(b)$ is a r.v. whose value depends on the values of the ξ_i's.) Then there exists an absolute constant t_0 such that for all $t \geqslant t_0$, for all $j \in \{0, 1, ..., b\}$,

$$\Pr(N_\theta(b) = j) \leqslant \binom{b}{j}\left(\frac{1}{2}\right)^b \left(1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1 - t^{-1/3})\right)$$
$$\text{for} \quad 0 \leqslant j < b/2, \tag{5.90}$$

$$\Pr(N_\theta(b) = j) \geqslant \binom{b}{j}\left(\frac{1}{2}\right)^b \left(1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1 - t^{-1/3})\right)$$
$$\text{for} \quad b/2 < j \leqslant b. \tag{5.91}$$

Proof. Because of the exchangeability of the θ_i's, it will suffice to consider

$$q(j) := \Pr(\theta_1 = \cdots = \theta_j = +1, \theta_{j+1} = \cdots = \theta_b = -1).$$

Thus the lemma will follow if we can show that for all sufficiently large t,

$$q(j) \leqslant \left(\frac{1}{2}\right)^b \left(1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1 - t^{-1/3})\right)$$
$$\text{for} \quad 0 \leqslant j < b/2$$

and

$$q(j) \geqslant \left(\frac{1}{2}\right)^b \left(1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1 - t^{-1/3})\right)$$
$$\text{for} \quad b/2 < j \leqslant b. \tag{5.92}$$

We give the proof for the case $j > b/2$; the proof for $j < b/2$ is essentially the same. By symmetry, for all realizations $\hat{\theta}_1, ..., \hat{\theta}_t$ arising from realizations $\hat{\xi}_1, ..., \hat{\xi}_t$, we have

$$\begin{aligned}\Pr(\theta_1 = \hat{\theta}_1, ..., \theta_b = \hat{\theta}_b) &= \Pr(\theta_1 = \hat{\theta}_1, ..., \theta_b = \hat{\theta}_b \mid \xi_{maj} = 1) \\ &= 2\Pr(\xi_{maj} = 1, \xi_1 = \hat{\theta}_1, ..., \xi_b = \hat{\theta}_b).\end{aligned}$$

Define the event E_j by

$$E_j = \{\xi_1 = \cdots = \xi_j = 1, \xi_{j+1} = \cdots = \xi_b = -1\}. \tag{5.93}$$

Supposing from now on that $j > b/2$, we therefore have

$$\begin{aligned}q(j) &= 2\Pr\left(\left(\sum_{i=1}^t \xi_i > 0\right), E_j\right) \\ &= 2\Pr\left(E_j, \left(\sum_{i=b+1}^t \xi_i > b - 2j\right)\right) \\ &= 2^{-b+1}\Pr\left(\sum_{i=b+1}^t \xi_i > b - 2j\right) \\ &= 2^{-b+1}\left(\Pr\left(\sum_{i=b+1}^t \xi_i > 0\right)\right. \\ &\quad \left. + \Pr\left(b - 2j + 1 \leqslant \sum_{i=b+1}^t \xi_i \leqslant -1\right)\right) \\ &= 2^{-b+1}\left(\tfrac{1}{2} + 2^{b-t}\sum_{m=1}^{j-b/2}\binom{t-b}{(t-b+1)/2-m}\right),\end{aligned}$$

where the fourth equality in the third line uses $\sum_{i=b+1}^t \xi_i \neq 0$ (since $t-b$ is odd).

Therefore for $j > b/2$,

$$q(j) = 2^{-b}\left(1 + 2^{b-t+1}\sum_{m=1}^{j-b/2}\binom{t-b}{(t-b+1)/2-m}\right). \tag{5.94}$$

By the monotonicity of the binomial coefficients in the range of interest, we have for all terms in the sum

$$\begin{aligned}\binom{t-b}{(t-b+1)/2-m} &\geqslant \binom{t-b}{(t-b+1)/2-(j-b/2)} \\ &\geqslant \binom{t-b}{(t-b+1)/2-b/2}.\end{aligned}$$

Recalling that $b \leqslant t^{1/10}$ and using Stirling's formula, we find that

$$\binom{t-b}{(t-b-1)/2} = \sqrt{\frac{2}{\pi(t-b)}}\,2^{t-b}(1 + O(t^{-1})).$$

Now for $1 \leqslant m \leqslant j - b/2 - 1$, we have

$$\binom{t-b}{(t-b+1)/2-(m+1)} \Big/ \binom{t-b}{(t-b+1)/2-m} = \frac{(t-b+1)/2-m}{(t-b+1)/2+m} \geqslant 1 - t^{-4/5}.$$

It follows that (for all sufficiently large t), for all $m \in \{1, ..., j - b/2\}$,

$$\binom{t-b}{(t-b+1)/2-m} \geqslant \sqrt{\frac{2}{\pi t}}\,2^{t-b}(1 - t^{-1/3})$$

and that

$$q(j) \geqslant \left(\frac{1}{2}\right)^b \left(1 + 2\left(j - \frac{b}{2}\right)\sqrt{\frac{2}{\pi t}}(1 - t^{-1/3})\right), \tag{5.95}$$

as we wished to prove. ∎

LEMMA 5.20. *Let t, b, $\{\xi_i\}$, ξ_{maj}, $\{\theta_i\}$, etc., be defined as in Lemma* 5.19. *Let $\psi_1, ..., \psi_b$ be mutually i.i.d. Bernoulli r.v.'s with*

$$\psi_i = \begin{cases} +1 & w.p.\ \frac{1}{2} + (1-t^{-1/2})/\sqrt{2\pi t} \\ -1 & w.p.\ \frac{1}{2} - (1-t^{-1/8})/\sqrt{2\pi t}. \end{cases}$$

Let $N_\psi(b) = |\{i \in \{1, ..., b\} : \psi_i = 1\}|$. Then there exists an absolute constant t_0 such that for all $t \geqslant t_0$, for all $j \in \{0, ..., b\}$,

$$\Pr(N_\psi(b) = j) \geqslant \binom{b}{j}\left(\frac{1}{2}\right)^b \left(1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1-t^{-1/3})\right) \quad \text{for } 0 \leqslant j < b/2, \tag{5.96}$$

$$\Pr(N_\psi(b) = j) \leqslant \binom{b}{j}\left(\frac{1}{2}\right)^b \left(1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1-t^{-1/3})\right) \quad \text{for } b/2 < j \leqslant b. \tag{5.97}$$

Proof. As in Lemma 5.19, we are interested in two cases, but we prove only one of the two, since the proof of the other case is essentially the same.

$$\begin{aligned} \Pr(N_\psi(b) = j) &= \binom{b}{j}\left(\frac{1}{2} + \frac{1}{\sqrt{2\pi t}}(1-t^{-1/8})\right)^j \\ &\quad \times \left(\frac{1}{2} - \frac{1}{\sqrt{2\pi t}}(1-t^{-1/8})\right)^{b-j} \\ &= \binom{b}{j}\left(\frac{1}{2}\right)^b \left(1 + \sqrt{\frac{2}{\pi t}}(1-t^{-1/8})\right)^j \\ &\quad \times \left(1 - \sqrt{\frac{2}{\pi t}}(1-t^{-1/8})\right)^{b-j}. \end{aligned}$$

Recalling that $j \leqslant b \leqslant t^{1/10}$ and letting

$$\gamma := (1-t^{-1/8})\sqrt{\frac{2}{\pi t}},$$

we have

$$\Pr(N_\psi(b) = j) = 2^{-b}\binom{b}{j}(1+\gamma)^j(1-\gamma)^{b-j}. \tag{5.98}$$

Choose any ϵ such that $0 < \epsilon < 1/100$. Then

$$\begin{aligned} (1+\gamma)^j(1-\gamma)^{b-j} &= (1+\gamma)^{2j-b}\left[(1+\gamma)^{b-j}(1-\gamma)^{b-j}\right] \\ &= (1+\gamma)^{2j-b}(1-\gamma^2)^{b-j}, \end{aligned}$$

and it is easy to see that for all sufficiently large t,

$$1 - t^{-2/5+\epsilon} \leqslant (1-\gamma^2)^{b-j} \leqslant 1 \quad \text{for all } j \in \{0, ..., b\}.$$

Applying the binomial formula to $(1+\gamma)^{2j-b}$, for all sufficiently large t, we easily obtain

$$\begin{aligned} &1 + (2j-b)\,\gamma(1-t^{-3/10+\epsilon}) \\ &\quad \leqslant (1+\gamma)^{2j-b} \\ &\quad \leqslant 1 + (2j-b)\,\gamma(1+t^{-3/10+\epsilon}) \qquad \text{for all } \ j > b/2, \end{aligned}$$

and a similar result holds for $j < b/2$. It follows that for all sufficiently large t, uniformly for $j > b/2$,

$$(1+\gamma)^{2j-b}(1-\gamma^2)^{b-j} \leqslant 1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1-t^{-1/8-\epsilon}) \tag{5.99}$$

$$\leqslant 1 + (2j-b)\sqrt{\frac{2}{\pi t}}(1-t^{-1/3}). \tag{5.100}$$

Combining (5.98) and (5.99), we obtain (5.97). Inequality (5.96) follows similarly. ∎

LEMMA 5.21. *Let t, b, $\{\xi_i\}$, $\{\theta_i\}$, etc., be defined as in the previous two lemmas. Then for all sufficiently large t (uniformly for all b and j),*

$$\Pr(N_\theta(b) \leqslant j) \leqslant \Pr(N_\psi(b) \leqslant j) \qquad \text{for all } \ j \in \{0, 1, ..., b\}. \tag{5.101}$$

Proof. Comparing the two inequalities (5.90) and (5.91) in Lemma 5.19 with the corresponding inequalities (5.96) and (5.97) in Lemma 5.20, we see that for all sufficiently large t, for $0 \leqslant i \leqslant b$,

$$\Pr(N_\theta(b) = i) \leqslant \Pr(N_\psi(b) = i), \qquad \forall i < b/2 \tag{5.102}$$

and

$$\Pr(N_\theta(b) = i) \geqslant \Pr(N_\psi(b) = i), \qquad \forall i > b/2. \tag{5.103}$$

Summing (5.102) from $i = 0$ to j, we have

$$\Pr(N_\theta(b) \leqslant j) \leqslant \Pr(N_\psi(b) \leqslant j), \quad \forall j < \frac{b}{2}. \tag{5.104}$$

Summing (5.103) from $i = j$ to b, we have

$$\Pr(N_\theta(b) \geqslant j) \geqslant \Pr(N_\psi(b) \geqslant j), \qquad \forall j > b/2.$$

Equivalently,

$$\Pr(N_\theta(b) \leqslant j-1) \leqslant \Pr(N_\psi(b) \leqslant j-1), \qquad \forall j > b/2. \tag{5.105}$$

Combining (5.104) and (5.105), together with the fact that

$$\Pr(N_\theta(b) \leqslant b) = \Pr(N_\psi(b) \leqslant b) = 1,$$

yields (5.101). ∎

Lemma 5.1 will now follow immediately from the following lemma.

LEMMA 5.22. *Suppose that* $\theta_1, ..., \theta_m$ *are exchangeable* $0-1$ *r.v.'s, as are* $\psi_1, ..., \psi_m$. *Let* $u := \sum_{i=1}^m \theta_i$ *have c.d.f.* $F_u(\cdot)$, *let* $v := \sum_{i=1}^m \psi_i$ *have c.d.f.* $F_v(\cdot)$, *and suppose that* $F_u \leqslant F_v$.

Then we can define $\psi'_1, ..., \psi'_m$ *on the same space as the* θ*'s, where* $\psi'_1, ..., \psi'_m$ *have the same joint distribution as* $\psi_1, ..., \psi_m$, *and where*

$$\psi'_i \leqslant \theta_i, \quad \forall i \ \ (w.p.1).$$

Proof. Since the θ's are exchangeable, we can think of generating them by the following three-step procedure.

Step 1: Choose the real auxiliary r.v. x uniformly from $(0, 1)$, and set $u = \hat{u} \in \{0, 1, ..., m\}$, where $F_u(\hat{u}-1) < x \leqslant F_u(\hat{u})$.

Step 2: Set $\tilde{\theta}_1 = \cdots = \tilde{\theta}_{\hat{u}} = 1$, $\tilde{\theta}_{\hat{u}+1} = \cdots = \tilde{\theta}_m = 0$.

Step 3: Let $\theta_1, ..., \theta_m$ be a random permutation of the $\tilde{\theta}$'s:

$$\theta_i = \tilde{\theta}_{\pi(i)}, \quad \forall i, \qquad \text{for} \quad \pi \text{ random.}$$

We can generate the $\{\psi'_i\}$ in essentially the same way, choosing the realization $\hat{v}$ of the r.v. v according to the c.d.f. F_v. Then, by using the *same* r.v. x and random permutation π that we used for the $\{\theta_i\}$, we ensure that $\tilde{\psi}'_i \leqslant \tilde{\theta}_i$ and $\psi'_i \leqslant \theta_i \ \forall i$. ∎

5.5. *Proofs of Lemma* 5.3 *and Lemma* 5.4

DEFINITION. In the lemmas in this section, let $q := 1 - p$.

Proof of Lemma 5.3. Applying Lemma 5.2 to compute $b(m, p, \lceil mp + x \rceil)$, we find that for all sufficiently large m,

$$\begin{aligned} b(m, p, \lceil mp + x \rceil) &\leqslant \exp(-x^2/(2mpq)) \\ &\leqslant \exp(-(1/2)(mpq)^{2(2/3-\epsilon_1)-1}) \\ &\leqslant \exp(-(1/2)(m(m^{-1/3+\epsilon_1})(\tfrac{1}{2}))^{1/3-2\epsilon_1}) \\ &= \exp(-(1/2)(\tfrac{1}{2}m^{2/3+\epsilon_1})^{1/3-2\epsilon_1}) \\ &\leqslant \exp(-(m^{2/3+\epsilon_1/2})^{1/3-2\epsilon_1}). \end{aligned}$$

But now $\forall j \geqslant mp + x$, either $j + 1 \geqslant m$ (in which case $b(m, p, j+1) = 0$) or $x < m(1-p)$, in which case

$$\begin{aligned} \frac{b(m, p, j+1)}{b(m, p, j)} &= \frac{m-j}{j+1}\frac{p}{q} \leqslant \frac{m-(mp+x)}{mp+x}\frac{p}{q} \\ &= \frac{mq-x}{mp+x}\frac{p}{q} = \frac{1-x/mq}{1+x/mp} \\ &= \frac{1-px/mpq}{1+qx/mpq} \leqslant \frac{1}{(1+px/mpq)(1+qx/mpq)} \\ &= \frac{1}{1+x/mpq+x^2/m^2pq} \leqslant \frac{1}{1+x/mpq} \\ &\leqslant \frac{1}{1+(mpq)^{-(1/3+\epsilon_1)}} \leqslant \frac{1}{1+(mpq)^{-1/2}} \\ &\leqslant \frac{1}{1+m^{-1/2}} < 1 - \frac{1}{2}m^{-1/2}. \end{aligned}$$

Then by comparison with a geometric series, we find that

$$\begin{aligned} \sum_{j \geqslant mp+x} b(m, p, j) &\leqslant 2m^{1/2} b(m, p, \lceil mp + x \rceil) \\ &\leqslant \exp(-(m^{2/3})^{1/3-2\epsilon_1}) \\ &= \exp(-m^{(4/3)(1/6-\epsilon_1)}), \end{aligned}$$

as we wished to show. The bound for $\sum_{j \leqslant mp-x} b(m, p, j)$ follows in exactly the same way if we switch the roles of p and q. ∎

Proof of Lemma 5.4. We wish to show that $\forall \epsilon > 0$, $\exists m_0 = m_0(\epsilon)$ (not depending on p or j) such that $\forall m \geqslant m_0$, $\forall j \in \{0, 1, ..., m\}$, $\forall p \in [1/10, 9/10]$,

$$B(m, p, j) \leqslant \exp(-m^{1/6}) + (1+\epsilon)\, \Phi((j-mp)/(mpq)^{1/2}).$$

The second term on the right follows from Lemma 5.2 essentially as in the usual proof of the DeMoivre–Laplace theorem. The first (exponential) term on the right makes the lemma valid even when $|j - mp| \gg (mpq)^{1/2}$. The lemma follows easily from Lemma 5.2 and Lemma 5.3. ∎

5.6. *Proof of Lemma* 5.5

The main idea used in the proof of Lemma 5.5 is the generalized Markov inequality given in the following lemma.

LEMMA 5.23. *Suppose that* $\xi_1, ..., \xi_t$ *are exchangeable indicator r.v.'s* (0–1 *r.v.'s*), *and suppose that* $1 \leqslant b < t$.

Let $P_b := \Pr(\xi_1 = \xi_2 = \cdots = \xi_b = 1)$. *Then for any integer* $T \in [b, t-1]$,

$$\Pr\left(\sum_{i=1}^t \xi_i \geqslant T\right) \leqslant \frac{\binom{t}{b} P_b}{\binom{T}{b}}.$$

Proof. Note that

$$\mathrm{E}\left[\sum_{\substack{B\subseteq[t],\\ |B|=b}} I\{\xi_i=1\ \forall i\in B\}\right]=\binom{t}{b}\Pr(\xi_i=1\ \forall i\in B)$$

$$=\binom{t}{b}P_b.$$

If $\sum_{i=1}^{t}\xi_i\geqslant T\geqslant b$, then

$$\sum_{\substack{B\subseteq[t],\\ |B|=b}} I\{\xi_i=1\ \forall i\in B\}\geqslant\binom{T}{b}.$$

Therefore by Markov's inequality, for any $T\geqslant b$,

$$\Pr\left(\sum_{i=1}^{t}\xi_i\geqslant T\right)\leqslant\frac{\binom{t}{b}P_b}{\binom{T}{b}}.\quad\blacksquare$$

Using Lemma 5.23, we now prove Lemma 5.5 from Section 5.3.

CLAIM. *For* $b=2\lfloor(1/2)\,t^{1/10}\rfloor$ *and* $\Pr(\xi_1=\cdots=\xi_b=1)\leqslant q^b$, *there exists an absolute constant* t_0 *such that for all* $t\geqslant t_0$,

$$\Pr\left(\sum_{i=1}^{t}\xi_i\geqslant t^{3/5}+(1+t^{-1/12})\,qt\right)\leqslant\exp(-t^{1/70}).$$

Proof. Consider three cases, based on the value of q. All constants implied by $o(\ )$ and $O(\ )$ are independent of q and of the joint distribution of the $\{\xi_i\}$.

Case 1. Suppose first that

$$t^{-1/2}\leqslant q\leqslant\frac{1}{1+t^{-1/12}}.$$

Then

$$\Pr\left(\sum_{i=1}^{t}\xi_i\geqslant t^{3/5}+(1+t^{-1/12})\,qt\right)$$

$$\leqslant\Pr\left(\sum_{i=1}^{t}\xi_i\geqslant(1+t^{-1/12})\,qt\right)$$

$$\leqslant\frac{t(t-1)\cdots(t-(b-1))}{(cqt)(cqt-1)\cdots(cqt-(b-1))}\,q^b,$$

where $b=2\lfloor(\frac{1}{2})\,t^{1/10}\rfloor$ and $c=1+t^{-1/12}$. Continuing the chain of inequalities, we have

$$\cdots\leqslant\left(\frac{t-t^{1/10}}{cqt-t^{1/10}}\right)^b q^b$$

$$=\left(\frac{qt-qt^{1/10}}{qt-(1/c)\,t^{1/10}}\right)^b\left(\frac{1}{c}\right)^b$$

$$=\left(\frac{1}{c}\right)^b(1+O(t^{-3/10}))$$

$$=\exp(-b\log c+o(1))$$

$$=\exp(-(t^{1/10})(t^{-1/12})(1+o(1))+o(1))$$

$$=\exp(-t^{1/60}(1+o(1)))\ll\exp(-t^{1/70}),$$

as desired.

Case 2. Suppose that

$$\frac{1}{1+t^{-1/12}}\leqslant q\leqslant 1.$$

Then clearly $\Pr(\sum_{i=1}^{t}\xi_i\geqslant t^{3/5}+(1+t^{-1/12})\,qt)=0$.

Case 3 (final case). $0\leqslant q\leqslant t^{-1/2}$. Then

$$\Pr\left(\sum_{i=1}^{t}\xi_i\geqslant t^{3/5}+(1+t^{-1/12})\,qt\right)$$

$$\leqslant\Pr\left(\sum_{i=1}^{t}\xi_i\geqslant\lceil t^{3/5}\rceil\right)$$

$$\leqslant\frac{t(t-1)\cdots(t-(b-1))}{(t^{3/5})(t^{3/5}-1)\cdots(t^{3/5}-(b-1))}\,q^b$$

$$\leqslant\left(\frac{t-t^{1/10}}{t^{3/5}-t^{1/10}}\,q\right)^b\leqslant\left(\frac{t^{0.5}-t^{-0.4}}{t^{0.6}-t^{0.1}}\right)^b$$

$$=\exp\{-(t^{0.1}+o(1))\log(t^{0.1}(1+o(1)))\}$$

$$\ll\exp(-t^{1/70}).\quad\blacksquare$$

5.7. *Proof of Lemma* 5.7

CLAIM. *Given a collection of* t *real numbers* $c_1\leqslant\cdots\leqslant c_t$ (*not necessarily distinct*), *define the corresponding empirical c.d.f.* $\hat{F}$ *in the usual way. Define mutually i.i.d. r.v.'s* $\xi_1,\ldots,\xi_t$, *each with statistical c.d.f.*

$$G(u)=\begin{cases}\min\{t^{-1/6}+\hat{F}(u),\,1\} & \text{if}\quad u\geqslant c_1\\ 0 & \text{if}\quad u<c_1,\end{cases}$$

and let $\hat{G}(\cdot)$ *be the* sample *c.d.f. of the* $\{\xi_i\}$. *Then there is an absolute constant* t_0 *such that for all* $t\geqslant t_0$,

$$\Pr(\hat{G}(u)\geqslant\hat{F}(u)\ \forall u\in\mathbb{R})\geqslant 1-\exp(-t^{1/9}).$$

Proof. The function $\hat{F}$ is piecewise constant with jumps of height $1/t$ (or multiples thereof). Let $\xi^{(1)}\leqslant\cdots\leqslant\xi^{(t)}$ be

the order statistics of $\xi_1, ..., \xi_t$, with $\hat{\xi}^{(1)} \leqslant \cdots \leqslant \hat{\xi}^{(t)}$ their sample values. Since $\hat{G}$ is nondecreasing and $\hat{F}$ is piecewise constant, it will be enough for us to show that for all $t \geqslant t_0$,

$$\Pr(\hat{G}(c_j) \geqslant j/t \qquad \forall j \in [t]) \geqslant 1 - \exp(-t^{1/9}).$$

It will therefore suffice to show that $\forall t \geqslant t_0$, $\forall j \in [t]$,

$$\Pr\left(\hat{G}(c_j) < \frac{j}{t}\right) \leqslant \frac{1}{t}\exp(-t^{1/9}).$$

Now

$$\begin{aligned}\Pr&\left(\hat{G}(c_j) < \frac{j}{t}\right)\\ &= \Pr(\exists \text{ at most } j-1 \text{ different indices } i \text{ s.t. } \xi_i \leqslant c_j)\\ &= \sum_{0 \leqslant l < j} \binom{t}{l} p^l (1-p)^{t-l},\end{aligned}$$

where for $j \geqslant 1$,

$$p := \Pr(\xi_1 \leqslant c_j) = G(c_j) = \min\left\{\frac{j}{t} + t^{-1/6}, 1\right\}.$$

If $j \geqslant t - t^{5/6}$, then $p = 1$ and it follows that

$$\Pr\left(\hat{G}(c_j) < \frac{j}{t}\right) = 0 \leqslant \frac{1}{t}\exp(-t^{1/9}),$$

as desired.

Now assume that $1 \leqslant j < t - t^{5/6}$, so that

$$p = \frac{j}{t} + t^{-1/6} < 1.$$

Letting

$$D(p) := \log(p^l (1-p)^{t-l}),$$

we have

$$D'(p) = \frac{l}{p} - \frac{t-l}{1-p} = \frac{l - pt}{p(1-p)}.$$

Thus $D'(p) < 0$ for $l < j < pt$, so $D(p)$ is a decreasing function of p. Therefore a lower bound on p yields an upper bound on the sum:

$$\Pr(\hat{G}(c_j) < j/t) \leqslant \sum_{0 \leqslant l < j} \binom{t}{l} (p^*)^l (1-p^*)^{t-l},$$

where $p^* = j/t + (\frac{1}{2})\, t^{-1/6}$. Then for $j < t - t^{5/6}$, we have $|j - tp^*| = \frac{1}{2}t^{5/6} \geqslant t^{2/3}$ and $\frac{1}{2}t^{-1/6} \leqslant p^* \leqslant 1 - \frac{1}{2}t^{-1/6}$, so Lemma 5.3 applies with $\epsilon_1 = 1/100$. Thus for all t sufficiently large, we have (uniformly in p^* and j)

$$\begin{aligned}\sum_{l < j} \binom{t}{l} (p^*)^l (1-p^*)^{t-l} &\leqslant \exp(-t^{4/3(1/6 - 1/100)})\\ &\ll \exp(-t^{1/9}). \quad \blacksquare\end{aligned}$$

ACKNOWLEDGMENTS

We thank H. S. Seung and L. Shepp for communicating this problem and for helpful discussions. We also thank S. Venkatesh and an anonymous reviewer for bringing some related work to our attention.

REFERENCES

1. N. Alon and J. Spencer, "The Probabilistic method," Wiley, New York, 1992.
2. E. Barkai and I. Kanter, Storage capacity of a multilayer neural network with binary weights, *Europhys. Lett.* **14** (1991), 107–112.
3. B. Bollobás, "Random Graphs," Academic Press, London, 1985.
4. L. Breiman, "Probability," Addison–Wesley, Reading, MA, 1968.
5. T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electronic Comput.* **14** (1965), 326–331.
6. S. Fang and S. Venkatesh, On the average tractability of binary integer programming and the curious transition to perfect generalization in learning majority functions, *in* "Proceedings of the 6th Workshop on Computational Learning Theory," Morgan Kaufman, San Mateo, CA, 1993.
7. S. Fang and S. Venkatesh, The capacity of majority rule, *Random Structures and Algorithms* **12** (1998), 83–109.
8. S. Fang and S. Venkatesh, Learning binary perceptrons perfectly efficiently, *J. Comput. System Sci.* **52** (1996), 374–389.
9. Z. Füredi, Random polytopes in the *d*-dimensional cube, *Discrete Comput. Geom.* **1** (1986), 315–319.
10. E. Gardner, Maximum storage capacity in neural networks, *Europhys. Lett.* **4** (1987), 481–485.
11. E. Gardner, The space of interactions in neural network models, *J. Phys. A* **21** (1988), 257–270.
12. E. Gardner and B. Derrida, Optimal storage properties of neural network models, *J. Phys. A* **21** (1988), 271–284.
13. M. Golea and M. Marchand, Average case analysis of the clipped Hebb rule for nonoverlapping perceptron networks, *in* "Proceedings of the 6th Workshop on Computational Learning Theory," pp. 151–157, Morgan Kaufman, San Mateo, CA, 1993.
14. G. Gyorgyi, First-order transition to perfect generalization in a neural network with binary synapses, *Phys. Rev. A* **41** (1990), 7097–7100.
15. D. Haussler, M. Kearns, H. S. Seung, and N. Tishby, Rigorous learning curve bounds from statistical mechanics, *in* "Proceedings of the 7th Workshop on Computational Learning Theory," pp. 76–87, Morgan Kaufman, San Mateo, CA, 1994.
16. D. O. Hebb, "The Organization of Behavior," Wiley, New York, 1949.
17. G. E. Hinton and J. A. Anderson (Eds.), "Parallel Models of Associative Memory," Erlbaum, Hillsdale, NJ, 1981.

18. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci. U.S.A.* **79** (1982), 2554–2558.
19. J. Kahn, J. Komlós, and E. Szemerédi, Singularity probability for random ± 1 matrices, preprint, 1993.
20. H. Köhler, S. Diederich, W. Kinzel, and M. Opper, Learning algorithm for a neural network with binary synapses, *J. Phys. B* **78** (1990), 333–342.
21. T. Kohonen, "Self-Organization and Association Memory," Springer-Verlag, New York, 1984.
22. T. Kohonen, State of the art in neural computing, *in* "IEEE First International Conference on Neural Networks I," pp. 77–90, 1987.
23. J. Komlós, On the determinant of (0, 1) matrices, *Studia Sci. Math. Hung.* **2** (1967), 7–21.
24. J. Komlós and R. Paturi, Convergence results in an associative memory model, *Neural Networks* **1** (1988), 239–250.
25. W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, *J. Physique* **50** (1989), 3057–3066.
26. W. Krauth and M. Opper, Critical storage capacity of the $j = \pm 1$ neural network, *J. Phys. A*: *Math. Gen.* **22** (1989), L519–L523.
27. Y. S. Lashley, "The Neurophysiology of Lashley: Selected Papers of K. S. Lashley," McGraw–Hill, New York, 1960.
28. W. A. Little, The existence of persistent states in the brain, *Math. Biosci.* **19** (1974), 101–119.
29. W. A. Little and G. L. Shaw, Analytic study of the memory storage capacity of a neural network, *Math. Biosci.* **39** (1978), 281–290.
30. G. Mato and N. Parga, Generalization properties of multilayered neural networks, *J. Phys. A*: *Math. Gen.* **25** (1992), 5047–5054.
31. W. S. McCulloch and W. H. Pitts, A logical calculus for ideas immanent in nervous activity, *Bull. Math. Biophys.* **5** (1943), 115–133.
32. P. Moran, "An Introduction to Probability Theory," Oxford Univ. Press, Oxford, UK, 1968.
33. H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, *Phys. Rev. A* **45** (1992), 6056–6091.
34. S. Venkatesh, Computation and learning in the context of neural network capacity, *in* "Neural Networks for Perception" (H. Wechsler, Ed.), Academic Press, New York, 1991.
35. S. Venkatesh, On learning binary weights for majority functions, *in* "Proceedings of the 4th Workshop on Computational Learning Theory," pp. 257–266, Morgan Kaufman, San Mateo, CA, 1991.
36. S. Venkatesh, The science of making ERORS: What error tolerance implies for capacity in neural networks, *IEEE Trans. Knowledge and Data Eng.* **4** (1992), 135–144.
37. J. Wendel, A problem in geometric probability, *Math. Scand.* **11** (1962), 109–111.
38. J. M. Wozencraft and I. M. Jacobs, "Principles of Communication Engineering," Wiley, New York, 1965.

Printed in Belgium